MEASURING PRE-SPEECH ARTICULATION

PERTTI PALO

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

QUEEN MARGARET UNIVERSITY

2019

Abstract: What do speakers do when they start to talk? This thesis concentrates on the articulatory aspects of this problem, and offers methodological and experimental results on tongue movement, captured using Ultrasound Tongue Imaging (UTI).

Speech initiation occurs at the start of every utterance. An understanding of the timing relationship between articulatory initiation (which occurs first) and acoustic initiation (that is, the start of audible speech) has implications for speech production theories, the methodological design and interpretation of speech production experiments, and clinical studies of speech production.

Two novel automated techniques for detecting articulatory onsets in UTI data were developed based on Euclidean distance. The methods are verified against manually annotated data. The latter technique is based on a novel way of identifying the region of the tongue that is first to initiate movement.

Data from three speech production experiments are analysed in this thesis. The first experiment is picture naming recorded with UTI and is used to explore behavioural variation at the beginning of an utterance, and to test and develop analysis tools for articulatory data.

The second experiment also uses UTI recordings, but it is specifically designed to exclude any pre-speech movements of the articulators which are not directly related to the linguistic content of the utterance itself (that is, which are not expected to be present in every full repetition of the utterance), in order to study undisturbed speech initiation. The materials systematically varied the phonetic onsets of the monosyllabic target words, and the vowel nucleus. They also provided an acoustic measure of the duration of the syllable rhyme. Statistical models analysed the timing relationships of articulatory onset, and acoustic durations of the sound segments, and the acoustic duration of the rhyme.

Finally, to test a discrepancy between the results of the second UTI experiment and findings in the literature, based on data recorded with Electromagnetic Articulography (EMA), a third experiment measured a single speaker using both methods and matched materials.

Using the global Pixel Difference and Scanline-based Pixel Difference analysis methods developed and verified in the first half of the thesis, the main experimental findings were as follows. First, pre-utterance silent articulation is timed in inverse correlation with the acoustic duration of the onset consonant and in positive correlation with the acoustic rhyme of the first word. Because of the latter correlation, it should be considered part of the first word. Second, comparison of UTI and EMA failed to replicate the discrepancy. Instead, EMA was found to produce longer reaction times independent of utterance type.

Keywords: Speech initiation, pre-speech articulation, delayed naming, ultrasound tongue imaging, electromagnetic articulography, automated methods.

To my parents
without whom this would not be.

Acknowledgements

This thesis has been written for Queen Margaret University, Speech and Hearing Sciences. Funding has been received from Queen Margaret University in the form of a three-year bursary, Emil Aaltonen Foundation (6 month working grant), and my parents Helena and Veikko Palo.

I want to thank my supervisors Dr. Sonja Schaeffler, Prof. James M. Scobbie, Dr. Korin Richmond and my external advisor Dr. Juraj Šimko. Sonja and Jim have been the best supervisors that I have had the fortune to work with. They have provided advice, comments and encouragement to keep me going. Juraj has been especially helpful with Mona Lehtinen in making it possible to run the EMA experiment in Helsinki and also providing aid in the form of post-processing the data from that experiment.

On the note of help with experiments, Prof. Alan Wrench and Steve Cowen have my heartfelt thanks. Alan has provided invaluable help in understanding the workings of AAA and in tirelessly fixing things when something has gone wrong with that software. As for Steve, I learned how to run an ultrasound experiment by emulating him. Both have also proven to be good friends.

I would also like to express my gratitude for my examiners, Dr. Susanne Fuchs and Dr. Robin Lickley, for their helpful comments on improving this thesis.

A special thank you is also due to many of my colleagues. This includes Pat, Eleanor, Felix, Ben, Stephen, Anna, Maria, and many others who have provided conversation and companionship which has made me feel like part of a larger community. And not to forget, I have had second academic homes in the UK at Universities of Strathclyde, Glasgow and Newcastle, where a lot of good chatting has happened with many people: My thanks go to Joanne, Susie, Jane, Ewa, Fabienne, Robert, Jalal, Ghada, and Danielle, to give a short and very incomplete list. And obviously also people further afield: My thanks to Scott, at least two Johns, Daniel, Martti, Tommi, Mike, Stina, Misha, and many more.

I would like to thank Kevin Roon for several discussions, but in particular for the one in July of 2014 that resulted in me selecting the materials in Ex-

periment 2 to be phonetically motivated rather than psychologically motivated. Both questions would have been interesting, but the one I chose is the one that I feel to be more so. Besides this rather concrete point to be grateful for, he also happens to be very insightful, encouraging and enthusiastic. Getting to talk to such people is always good for one's own motivation.

During my time in Edinburgh I have made a number of new friends who have provided very welcome relief from working on the thesis mainly in the form of music and dancing but also volunteering for the National Trust for Scotland. Nigel Gatherer was my first music teacher in Edinburgh, and Rebecca Knorr was the second one: Your teaching gave my music wings and I used them to flap around like a new hatchling. And not to forget Samuli Karjalainen who has been teaching me the last few years. His method of "cup of coffee and a chat with music" is a surprisingly efficient way of teaching.

Through Nigel's class I met many of my musical friends. Especially the people who on Wednesdays play in Leslie's bar – most notably Kate who encouraged me to join in with the playing when I was feeling a bit timid, and Jane who provided encouragement with getting the thesis done. Not to forget the good people of Edinburgh New Scotland Scottish Country Dancing Society, Edinbal, and Glasbal: Emilia, Kieran, J-C, Xavier, Minna, Emma, Isla, Lewis, Erin, two Davids, Greg, Edward, Rachel, and so many more. Without you I would not be a dancer.

In Edinburgh, I also had the good fortune of living close by to Douglas and Jane-Anne who arrange a lot of concerts at their house and other venues. Being able to visit their living room for evening gigs kept my spirits up while drudging through thesis work. Through them I also got to know a bunch of folky people, including Andy a.k.a. DJ Dolphin Boy, whose tracks have been an absolute treat.

Over the years I got to know some folks who live in pretty places and have been more than hospitable when I have come around. Andrew, Paul, Will, Becks, Hebe, Jack, and the good folks at the session at Forthingal: thank you kindly.

These last two years I have lived in Glasgow and had the good fortune to play roleplaying games with Dónal, Ian and two Micheals, as well as dance Morris dances with the Border Reivers – Dónal (yes, same one), Jess, Geoff, Niamh, one of the Michaels, Doug, Anna, Dani, Stuart, and a bunch of others. Ian gets a special mention for being one of the brothers I never had.

And then there are all the people back in Finland. Old friends from way back when, fellow Scouts and my Scout kids, friends from folk festivals at Kaustinen, Haapavesi, and Ruotsinpyhtää: More than two Mikkos, Jakke and his family, Pasi, at least two Anttis, Holtti, Santeri, Anu, Hanna, Annukka, Sara,

Aleksi, at least three Elinas, Jasmin, Heidi, Katri, V-P, Jani, Jonathan, Wasel, Marja, Klaus (and his banjo), Juho, Hessu, Jimmy, Sampo, Olli, and all the people that I can not recall right now. Kiitos.

Finally, my gratitude goes to my family: my parents, my sister and her family, and my girlfriend Caitlin. They have supported me through thick and thin. Thank you.

Espoo, September 2, 2019 Pertti Palo

Contents

List of Prior Publications of Thesis Material

Palo, P., Schaeffler, S., and Scobbie, J. M. (2014). Pre-speech tongue movements recorded with ultrasound. In *10th International Seminar on Speech Production (ISSP 2014)*, pages 304 – 307.

Palo, P., Schaeffler, S., and Scobbie, J. M. (2015a). Acoustic and articulatory speech reaction times with tongue ultrasound: What moves first? In *Ultrafest 2015*, Hong Kong.

Palo, P., Schaeffler, S., and Scobbie, J. M. (2015b). Effect of phonetic onset on acoustic and articulatory speech reaction times studied with tongue ultrasound. In *Proceedings of ICPhS 2015*, Glasgow, UK.

List of Figures

List of Tables

Symbols and Abbreviations

Symbols

/ɑ/ phoneme ɑ.

[ɑ] phone ɑ.

C Consonant phone or phoneme.

V Vowel phone or phoneme.

Abbreviations

2D Two Dimensional. 62, 65, 69,

3D Three Dimensional. 7, 34, 58, 60, 62, 64, 69, 70,

AAA Articulate Assistant Advanced. 103, 126, 127, 158, 162,

AAI Articulatory to Acoustic onset Interval. 54–57, 85–87, 145, 146, 149, 164, 165, 167, 169–172, 177, 178, 187, 188, 190, 192–194, 196, 198, 200–202, 206

EMA Electromagnetic Articulography. 7, 8, 10, 14, 55, 61–63, 69, 78, 79, 87, 89, 117, 148, 178–184, 186–190, 192, 194–198, 203–205

EMG Electromyography. 8, 25, 26, 56, 78, 104, 105, 110, 205

EPG Electropalatography. 7, 39, 58–61, 63, 78,

fps frames per second. 10, 65, 77, 98, 99, 122, 128, 158,

IPA International Phonetic Alphabet.

MRI Magnetic Resonance Imaging. 7, 8, 58, 63–67, 77, 79, 83, 92,

OD Onset consonant's acoustic Duration. 53, 85–87, 145–147, 153, 164, 167, 169, 171, 190, 192, 194–196, 199–201

OPG Optopalatography. 60, 61, 78,

PCA Principal Component Analysis. 81, 83,

PD Pixel Difference. 13, 83, 84, 88, 89, 92–107, 109–111, 114, 115, 119–122, 124, 125, 128, 129, 132, 133, 135–140, 142, 150, 201, 204–206, 227

SBPD Scanline Based Pixel Difference. 13, 88, 92, 104, 111, 113, 114, 117, 119–124, 163, 164, 177, 187, 196, 203, 205, 227

UTI Ultrasound Tongue Imaging. 1, 2, 7–10, 13, 14, 58, 70, 72, 74, 75, 77, 79–84, 87–89, 91, 92, 94, 96, 97, 99, 100, 114, 117, 125–127, 132, 145, 148–150, 158, 163, 177–180, 183, 184, 186–190, 192, 194–196, 198, 199, 203, 205

Chapter 1

Introduction

This thesis focuses on questions related to pre-speech movements, speech initiation, and the transition to speech from other modes of activity, such as mastication, swallowing – and most importantly in the experiments of this thesis – breathing quietly without articulatory movements. The two goals of this thesis are to analyse the timing relationships of articulatory activity of utterance initiation in relation to the acoustic activity of the utterance, and to develop methods for analysing speech initiation in Ultrasound Tongue Imaging (UTI) data.

The motivation to study the timing of pre-speech articulation and the way it relates to the acoustic timing of the utterance is to provide a baseline of timing of speech initiation that will be of use in future studies of both regular and pathological speech production. Guided by theories and results in the literature, and results derived from an experiment recorded before this work (Experiment 1), the main experiments (Experiments 2 and 3) have been designed to remove as many of the confounding factors (such as variations in the planning of an utterance and hesitations) as possible while keeping variation of relevant phonetic parameters (phonetic content of the beginning of the utterance, articulation rate) as part of the experiment. Results from these experiments further our understanding of the basic process of speech initiation.

To produce the results, new data analysis methods are developed in the thesis. Analysis of articulatory data – particularly ultrasound data – is both

difficult and time-consuming, because there are few tools available to aid in such analysis. What is particularly missing are visualisation tools that would display the evolution of the video over time in a single image without the need to flip through the video one frame at a time. This is comparable to analysing audio data without the aid of waveforms and spectrograms with the only option being listening to the samples over and over again. Neither are there any automatic time domain segmentation tools for UTI data. The method development in this thesis aims to address these needs.

The current chapter first looks at the rationale for studying specifically speech initiation (Section 1.1), then discusses using articulatory data in speech production research and specifically in a speech initiation project (Section 1.2), and the need to develop new analysis methods – including fully automated articulatory onset detection – as part of the project (Section 1.3). It then briefly discusses the materials and methods of the thesis in relation to the aims (Section 1.4), the operational definitions of *pre-speech* and *utterance* (Section 1.5), and finally, introduces each of the following chapters (Section 1.6).

1.1 Why speech initiation?

When we want to understand arm movements associated with throwing a ball in terms of efficiency and motor control, it is necessary to understand the physical movements of the throwing mechanism prior to and following the release of the ball. When we want to understand speech production in the same terms, it is necessary to understand the purposeful movements that lead to the utterance of speech both before and after the moment at which the first acoustic consequences of speech occur.

So, let us consider how speech starts. Which factors affect how speech starts and how do they affect it? There is a difference between producing speech and just articulating (while keeping mostly silent): In the former, speakers have the benefit of full auditory feedback, and in the latter they do not. Pre-speech articulation forms a special case as there is by definition no auditory speech,

and therefore, minimal auditory feedback as there is no acoustic speech signal (even though there may be breathing sounds, clicks produced by the opening vocal tract, and other non-speech sounds). At the same time the speaker is moving with the purpose of producing speech sounds. These silent articulatory stages vary depending on the initial sounds of the word or sentence about to be produced, but also because prior to speaking the vocal apparatus can be in a range of different modes such as breathing, masticating, or swallowing.

The phase of silent articulation is of course a special case, only found in the initiation of the first word in an utterance. This is because in continuous speech, the production of the beginning of each word overlaps with the end of the previous word. The sounds in the end of the previous word and the beginning of the next word are coarticulated with each other. Manipulating and observing experimentally the articulation in the silent phase at the beginning of a new utterance provides opportunities for theoretical explorations of speech production free from the usual coarticulatory effects of preceding speech segments.

Closely related to study of coarticulation, inter-speech postures and the hypothetical speech ready position – a position of the articulators that would be advantageous for initiating speech – have received attention in the recent years. Both are interesting from a speech motor control perspective and in how they may or may not be related to another hypothetical entity: the articulatory basis, which is a hypothesised neutral position of speech articulators, on which the articulatory targets of a language would be built upon.

Research into disfluencies, such as false starts and hesitations, and related pathologies is often already research into pre-speech, and so is research into inter-speech postures and the speech ready position, which are hypothesised in the literature. These behaviours are also potential practical application areas for results from such research. Basic speech initiation research can help in analysing and even correcting pathological behaviour patterns as well as essential theoretical information needed in implementing conversational systems that seek to mimic human behaviour.

From a methodological point of view, studying pre-speech challenges standard assumptions about results derived from speech reaction time studies and thus contributes to our understanding of theories of speech production. Many experimental paradigms rely on, or are confounded by, speech initiation: for example, speech reaction time experiments and speech experiments, where initiation is not studied, but the experiment consists of repeatedly initiating speech, such as reading one word at a time in random order. On the other hand, studies concentrate often on utterance medial phenomena based on the received wisdom that initial and final segments, syllables, and words always behave differently from those in the middle of an utterance. This is a sensible practice when we want to avoid effects due to f_0 declination, prosodic boundaries, final lengthening, and other such phenomena (Lehiste 1970, Oller 1973, Cho and Keating 2009). However, concentrating too much on the middle may well lead to theories that are ill-suited to account for beginnings and ends of utterances.

Speech initiation is ecologically relevant in conversation behaviour. After all, during a conversation, speakers will be continually initiating speech, speaking, stopping and listening only to start again – with all the richness of hesitations, reductions, false starts, signalling for turn taking, etc. Turn taking probably utilises also visual elements of speech and audible consequences of preparatory movements such as tongue clicks. However, measuring just the timing relationship of articulatory initiation and acoustic onset already gives us a baseline that will help in analysing turn taking behaviour and especially unsuccessful and abandoned attempts to inject a turn in conversation.

To summarise, studying speech initiation can shed light on how speech production is interleaved with other modes of action (breathing, mastication, swallowing, and so on). In the silent initial phases we can observe speech articulation free of carry over coarticulatory effects and of the effects of auditory feedback. Most importantly from the point of view of this thesis, studying timing of pre-speech articulation will give important information for studies of conversation behaviour of both regular and pathological speakers.

1.2 Why articulatory data?

Speech research generally studies the properties and characteristics of speech sounds, using a variety of acoustic analysis methods. The advantages of acoustic analysis include the central role that audible aspects of speech play in the transmission of spoken language (augmented by some visual elements), the ready availability of acoustic data from audio recordings of the voice, and the well-established analytic techniques employed in the acoustic analysis of speech. Acoustic analysis is sufficient for the majority of speech research because generating audible speech (which, by definition, is suitable for acoustic analysis) is the core function of speech production. When the interlocutor (or audience) can see the speaker's face the visible part of speech does improve speech comprehension (Sumby and Pollack 1954, Granström and House 2007). While speech is a multimodal phenomenon, the acoustic signal does remain the most important medium for transmitting linguistic information in speech, and the acoustic signal has been the focus of most research until recent times.

From the point of view of speech *production* research, there are some very important disadvantages in using only acoustic data. The most important is that acoustic data is only the end product of a very complex process – speech is a strong contender to be the most complex type of motor activity that humans engage in. As such, acoustic data provides only an indirect view of the process itself. In general, the process comprises cognitive processes leading to neural activation, respiratory changes and articulation – all of which combine to produce the main end product.

Fluent speech uses approximately 100 muscles (with about 100 separately controllable motor units per muscle) choreographed in a way that makes it possible to produce speech sounds at an average rate of about 15 per second. On top of the complexity of the motor control aspect of speech, there is also complexity in the articulatory-to-acoustic mapping, which is many-to-one. This means that more than one articulatory configuration can lead to the same – and especially perceptually the same – acoustic output. In other words, acoustic data

compresses the complexity of speech production into a single channel, whereas one of the central challenges of understanding the physical aspects of speech production is to account for its multiple and physiologically varied aspects. In particular, acoustic data provides no information at all on the silent but essential components of pre-speech and speech preparation.

The major and most complex articulator, used in the production of most consonants and central to the production of all vowels, is the tongue – consisting of a dozen muscles (four extrinsic and four paired intrinsic muscles). In speech production research, the movements of the tongue are a central topic of study. However, little attention is paid to how speakers transition from other modes of action (such as breathing, mastication and swallowing) to speech and how this is reflected in movements of the tongue. This thesis contributes to teasing apart the tongue's non-speech, pre-speech, and audible speech movements.

In summary, in speech production research using a combination of articulatory and acoustic data is preferable to using only the latter, because the latter is the end product of the speech production process while the former gives a more detailed view of the process itself. Furthermore, acoustic data compresses the complexity of the process in what in effect is lossy dimension reduction: more than one articulatory configuration may produce the same acoustic result. Finally, to better understand the way speakers initiate speech, we have to observe the silent articulation that leads to acoustic speech. Recording and analysing such data is the first aim of this thesis.

1.3 Why new analysis methods?

Writing on a project closely related to speech initiation, Roon (2013) (page 40) notes "Ideally, RT [reaction time] would be measured as the time between the presentation of the visual cue and the onset of articulatory movement associated with [the target utterance]." He goes on to state that the acquisition and analysis of articulatory reaction time data requires a prohibitive amount of time and is too expensive to be used in a project such as his – that is, a PhD thesis on

modelling the dynamics of phonological planning. What is needed instead is a project (probably several projects) where automation of articulatory analysis is a central part of the goals of the project.

Acquiring a large articulatory corpus is necessarily more expensive than acquiring a large acoustic corpus because the former requires more equipment. The extra cost can take the form of investing in infrastructure – when a new recording device such as an ultrasound machine or an Electromagnetic Articulography (EMA) device is bought, the form of paying for the use of a facility – such as paying for time at a Magnetic Resonance Imaging (MRI) machine, or the form of materials needed for running experiments – for example, continually replacing the EMA sensors or Electropalatography (EPG) palates as they wear out. However, this does not mean that there are no extensive articulatory corpora available (see for example, Turk et al. 2010) nor does it mean that articulatory data acquisition as such should be considered prohibitively expensive as there are relatively cheap methods available such as UTI or ordinary video recording of the lips and visible articulators.

Time-wise, the acquisition of articulatory data can be a bit slower or much slower than the acquisition of acoustic data. If recording static articulatory data, the acquisition time can be very long. Perhaps the extreme case is that of Three Dimensional (3D) MRI, where the acquisition of a single articulatory configuration can take several seconds or even minutes. Such methods, however, are not suitable for researching an essentially dynamic phenomenon like speech initiation. In contrast, recording dynamic articulatory data with methods such as UTI, EMA, or real-time MRI, which would potentially fit the needs of this thesis, the acquisition is at most only slightly slower than recording only acoustic data. In these cases depending on the articulatory recording method, the particular device, recording settings, and the computer hardware available, it might not be possible to save the data in real time, which can lead to longer inter-trial periods, while waiting for previous trial to be saved. But this is a mainly a limitation on how many recordings and how long recordings can be acquired in given time.

However, these problems get much worse with data analysis. There is

usually more than one articulatory signal being recorded (for example, multiple EMA sensors or UTI combined with a lip video recording). So, instead of labelling only the acoustic signal, with a combined articulatory acoustic corpus it is necessary to label both the acoustic signal and the articulatory signals. The acoustic data needs to be preferably segmentally and prosodically annotated so that – when for example dealing with speech initiation – turn initial phenomena can be isolated for study. Furthermore, articulatory data usually requires post processing whether it is in the form of signal enhancement (typical for EMA and Electromyography (EMG)) or dimension reduction (typical of UTI, MRI and other image based methods). All of this amounts to a prohibitively slow and costly analysis process. This then is the point where automation can step in to ease or completely eradicate the problem.

While studying speech initiation removes the confounding factor of carry-over coarticulation, other confounding factors are still present – such as anticipatory coarticulation and hesitations, to mention two. Thus, to properly understand the associated phenomena, studies need to analyse a substantial amount of articulatory and acoustic data. With current analysis tools such analysis is prohibitively time-consuming – the main bottleneck is in processing the articulatory data. In particular, UTI relies largely on manual annotation to detect the time points at which movement begins, or when maximum change occurs. Such annotation is very labour-intensive because unlike for audio data (which has waveforms and spectrograms) there are no *standard* tools for visualising change over an utterance in a single image, nor are there methods for automatic segmentation in the time domain.

Furthermore, the standard practice of tracing the midsagittal surface of the tongue at the air boundary discards a great deal of useful data from tongue-internal changes. We see that the amount of potentially useful data about the tongue in Figure 1.1 (a) is greatly reduced when it is simplified into Figure 1.1 (c) by tracing the tongue contour. This processing method leaves behind the rest of the anatomical data available in an ultrasound frame. In Figure 1.1 (a), we see anatomical structures beyond the tongue contour, such as the mandible

appearing as a shadow on the right, and the short tendon appearing as a brighter area to the left of the bottom of the mandible shadow. We also see a lot of noisy variation both above and below the tongue. This is usually called 'speckle noise' and, as we will see, is in fact produced by internal changes in the muscles and blood vessels that make up the tongue.

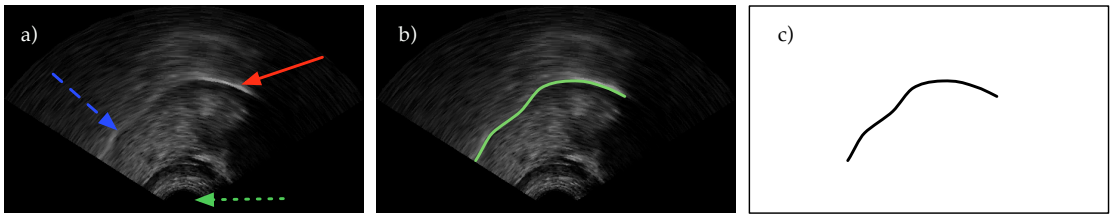

Figure 1.1: Extracting the tongue contour with a spline from a UTI frame. a) A UTI frame with arrows pointing to the top of the tongue surface near its tip (red, solid), tongue root (blue, long dashes), and the position of the ultrasound probe (green, short dashes), as well as some of the internal structure of the tongue visible, b) spline tracing of the tongue contour on the UTI frame (light green line), and c) the extracted spline (in black).

In conclusion, while articulatory data is necessary in studying speech initiation, it comes with the added cost of more time-consuming analysis than audio data and a lack of visualisation tools. In order to efficiently realise the first aim of this thesis – to record and analyse data of the silent articulation that leads to acoustic speech – these problems have to be addressed. To do so we need to develop automated methods that provide reliable timing information of articulatory events, and methods for visualising the changes in articulatory data, so that a human expert can verify results of automatic analysis. This is the second aim of this thesis.

1.4 Thesis overview

In this thesis, theoretical results are derived from a sequence of three interconnected articulatory experiments on speech initiation with adult participants. Ultrasound was chosen as the main recording method because it provides good

temporal resolution (better than 100 fps) combined with wide field of view (most of the mid-sagittal tongue outline) and richness of detail (many structures inside the tongue and other tissues are also visible in the data). It is also relatively easy on the participants.

All experiments of the thesis use Ultrasound Tongue Imaging (UTI) and the last experiment has a second part, which uses Electromagnetic Articulography (EMA). The data of the first experiment was mainly used to test new analysis tools in their early development stages. Its results were also crucial in deciding on which elements of pre-speech articulation to concentrate on. The second experiment answers the main theoretical research questions and the final (third) experiment addresses methodological issues rising from comparing results of the second experiment with results reported in literature.

As for data analysis, the labelling of the acoustic signals is nowadays a straight forward process with good tools available to aid in finding the relevant acoustic markers. The most important tools are the display of waveforms and spectrograms as functions of time, which makes it easy to observe the time evolution of the acoustic signal without having to listen to a sample repeatedly when trying to annotate acoustic boundaries. In contrast, there are no *standard* ways of displaying the relevant changes of ultrasound data as a function of time.

To facilitate answering the theoretical questions about timing, new analysis methods were developed for ultrasound data. An existing dimension reduction method – Euclidean distance of consecutive ultrasound frames – was first adapted for automated and manual onset detection and validated on a test set of ultrasound data from Experiment 2. It was then developed further to provide anatomically localised onset detection as well as a more robust method of automated onset detection.

To summarise, this thesis has two main goals: First, to analyse pre-speech tongue movements and relate their timing to the timecourse of the rest of the utterance. Second, to facilitate reaching the first goal, to develop automated methods for scientific analysis of Ultrasound Tongue Imaging data. To provide context for the transition from non-speech (waiting at rest in the experiments)

to speech, short, complete utterances were recorded and analysed in three articulatory experiments.

1.5 Definitions

In a wide sense, pre-speech articulation is articulation that takes place before the onset of acoustic speech. Based on this preliminary definition and the source filter theory of speech production (which will be discussed in Chapter 2) Figure 1.2 shows a conceptual timeline of speech initiation based on the source filter theory of speech production. The timeline represents both phenomena studied in this thesis (silent speech movements) and ones mainly excluded outside of its scope (preparatory movements).

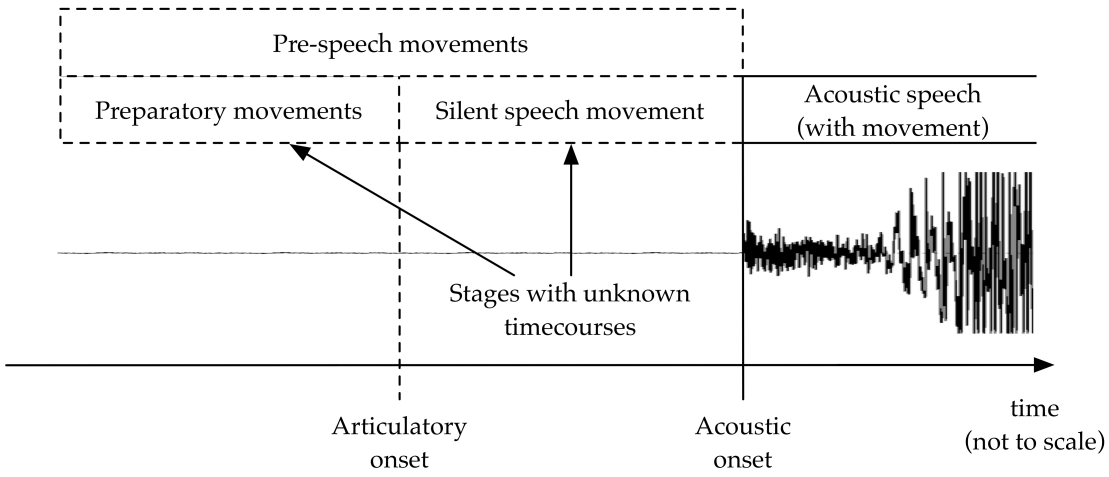

Figure 1.2: Conceptual timeline of pre-speech articulation and onset of acoustic speech. Pre-speech movements or pre-speech articulation consists of two stages: preparatory non-linguistic movements, which fall outside the scope of this thesis, and silent speech movement, which together with the timing of the acoustic utterance form the subject matter of this thesis.

Pauses in running speech, turn taking in interaction with an interlocutor, and phenomena in speeded trials or reaction time type tasks are included in this definition. However, non-speech movements of the articulators are specifically excluded from the definition of pre-speech used here and the scope of this thesis,

even if they are part of general non-linguistic preparation for speech. Outside of a laboratory it is not *always* possible to know which is which, but as we will see, this can be controlled with careful experimental design coupled with an understanding of how non-speech movements and speech movements can be differentiated in the processed data.

In non-experimental conditions it can be very difficult to clearly define what we mean by an utterance because the beginning and end of an utterance are frequently difficult to define due to hesitations before, during, and after speech, false starts, disfluencies, failed attempts to take a turn in a conversation, and other such natural speech phenomena.

In this thesis, pre-speech is defined as any articulatory movements, that take place in the mostly silent phase before the onset of acoustic output, and which are associated with speech articulation of the following utterance.

In this thesis, an utterance is defined as a piece of connected speech, which is preceded (and followed) by the speaker being silent. In the experiments of this thesis, an utterance is a single word spoken in this kind of isolation.

False starts and hesitations form a case which does not neatly fit in these definitions as the movements are not precisely preparatory in nature, but neither are they a part of the articulation of the first segments of the utterance. They are mainly excluded from the scope of this thesis except for the analysis of Experiment 1 in Chapter 4.

1.6 Structure of the thesis

This chapter and the following chapter (Chapters 1 and 2) provide the introduction and background of this thesis. They are followed by a methodological chapter (Chapter 3) describing the methods developed as part of this thesis. The experiments and their raw results can be found in Chapters 4–6. The final chapter (Chapter 7) provides overall results and discussion. Before moving on to the next chapter, we will look at the contents of each chapter in more detail below.

Chapter 2 reviews literature of both relevant speech production and initiation theories, as well as articulatory recording and analysis methods. A case is made for using Ultrasound Tongue Imaging as the main articulatory recording method in this thesis by comparing it to the best available alternative methods. Articulatory analysis methods are explored for the best candidate for further development in this thesis. Finally, theoretical research questions are formulated based on the theories and earlier research findings in the literature.

Chapter 3 describes the analysis methods developed in this thesis. It provides background on using Euclidean distance in analysing ultrasound and other video sequences and describes the development of novel analysis methods in this thesis beginning with basic Pixel Difference (PD) applied to raw ultrasound data, continuing with Scanline Based Pixel Difference (SBPD) to provide a localised change metric for ultrasound data, and two new automated activation onset detection methods. The chapter also gives validation results based on comparison of the new automated onset detection methods to manual analysis of Pixel Difference (PD) and manual video analysis.

The next three chapters describe the experiments, their data, and results with some initial discussion. Experiment 1 (in Chapter 4) is a picture naming experiment. Ultrasound data from the experiment was mainly used as a test bed in early method development and to explore the functioning of the new ultrasound data analysis methods.

Experiment 2 (Chapter 5) is a delayed naming experiment. A total of about 3700 tokens were recorded with ultrasound from four English-speaking participants. Unlike in Experiment 1, the phonetic content of the data was controlled to provide reasonable coverage of different English consonantal onsets. Analysis of ultrasound data from this experiment answers the main research questions about timing relations of silent initial articulation to the utterance being initiated.

Experiment 3 (Chapter 6) was motivated by a need to verify the results of Experiment 2. The need arose from an apparent discrepancy between results of Experiment 2 and a similar experiment in the literature. The purpose of

Experiment 3 is to check if the discrepancy could be explained by methodological differences – the experiments in question differ in recording methods. A single speaker (the author) was recorded in both EMA and Ultrasound Tongue Imaging (UTI). The results do not provide evidence for the discrepancy being due to differences between EMA and UTI, but do show a level difference in both articulatory and acoustic reaction times between EMA and UTI data. Since the experiment used only one speaker, in final analysis, the result is inconclusive.

Chapter 7 first answers the research questions based on the results of the experiments and provides critical discussion on the reliability of the results. The chapter then discusses the analysis methods developed in this thesis. Finally, it provides discussion on the wider implications of the results and possible future directions of research.

Chapter 2

Speech production, speech initiation and research methodology

This thesis is concerned with the interplay of acoustic and articulatory phenomena that precede or coincide with the beginning of acoustic speech production. This is a distinct but regularly occurring phase in running speech, that can be observed every time a speaker initiates a monologue or a conversation, begins a new turn in conversation, or continues after a pause in a monologue or conversation. As we will see, it is an understudied phenomenon. There is especially a lack of studies that would combine acoustic measurements with articulatory ones.

This chapter will review the relevant theories and related previous research as theoretical groundwork for the method development, the speech experiments, and the results presented in this thesis. The first section provides a general overview of the process of speaking taking into account anatomical structures involved in the process as well as acoustic, articulatory and psycholinguistic theories of speech production. It is argued that the interface between psycholinguistic/neurological theories and acoustic/articulatory theories deserves more attention and that speech initiation studies are well suited for studying this interface.

The second section concentrates on speech initiation studies and related

concepts, with particular emphasis on reaction time studies. We will conclude that articulatory measurements are needed to study speech initiation.

The third section provides an overview of the major articulatory recording methods, and makes a case for using tongue ultrasound imaging in this thesis. The fourth section addresses the issue of articulatory analysis by reviewing available analysis methods. It concludes that new methods are needed.

The final section introduces the main research questions and corresponding hypothesis.

2.1 Speech production

In order to understand speech initiation we need to understand its relation to speech production in general. We will consider speech as a process where different neural and physical stages follow onto each other and overlap with each other in time.

For purposes of illustrating the divide between theories and models of neural processes and those of physical processes Figure 2.1 shows a novel simplified synthesis of the models reviewed below (See Fant 1960, Mermelstein 1973, Sternberg et al. 1988, Levelt et al. 1999, Guenther et al. 2006, as well as Sections 2.1.6 and 2.1.7 below).

We see that Figure 2.1 divides speech production into a planning stage, a motor command generation stage, an execution stage, an articulator movement

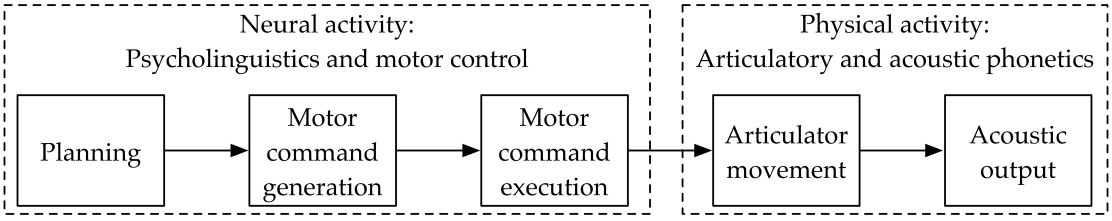

Figure 2.1: Main components of the speech production process. The arrows represent the flow of control and causation towards the goal of producing audible speech. This is a simplified view that excludes feedback loops from the representation.

stage, and an acoustic output stage. The first three stages are the concern of psycholinguistic theories and of speech motor control models. The last two stages are the concern of articulatory models and acoustic theories of speech production – in other words, of articulatory and acoustic phonetics. Before we review the most relevant theories and models, let us take a closer look at what role the stages play in the speech production process.

The first three stages belong to the domain of neural activity. In the planning stage thoughts are transformed into a linguistic form. What this form is considered to be depends on the theory in question. In the motor command stage neural commands are generated, which means an executable motor plan or program is assembled based on the linguistic representation. In the motor command execution stage, the current or predicted locations of the articulators are taken into account before sending neural signals that will trigger articulator movement.

Moving across the neural/physical activity divide, the remaining two stages belong to the domain of physical activity. The articulator movement stage is the result of the motor plan being executed, of neural control signals resulting in muscle activation. This includes respiratory activity and changes in muscle tone that do not necessarily cause movement, but that change the way an articulator physically responds to changes in air pressure and air flow through the vocal tract. Finally, the acoustic output stage is a consequence of articulator movements. The acoustic output consists of periodic and aperiodic energy of various frequencies and intensities – even silence when a complete closure of the vocal tract cuts of air flow and sound production – corresponding to the produced speech segments.

In running speech these stages overlap most of the time. While previously generated content is being produced as articulatory and acoustic output, new content is being planned and transformed to motor command sequences. It should be noted, that while the output has to be linear, the planning processes run in parallel (Indefrey and Levelt 2004, Guenther 2016).

The view presented so far is intentionally simplistic. The outline of stages

presented in Figure 2.1 is not meant to be rigid but instead should be taken as a tool for understanding the basic building blocks of most theories and models and how they relate to each other. Models and theories will often provide more refined definitions for each stage, with a breakdown into further stages or modules. It should also be noted that some theories and models will allow the order of stages to be changed or stages that overlap with others, for example in the context of modelling coarticulation or the coproduction of speech segments.

Aside from detail, the five stages above lack two important aspects of speech production: feedback loops and the fact that speech is meant to be heard by a listener. Both of these aspects will be briefly be reviewed later on, but as we will see, while important for speech production overall, they do not play as much of a role in speech initiation.

It is also worth noting that there are two pathways of tackling the problem of understanding speech production while bridging the neural/physical divide described above: one pathway working from the brain and central nervous system towards the acoustic speech signal and one pathway working from the acoustic speech signal towards the processes that produced it. In the following sections we will see that both approaches are necessary and that there is a gap between the two approaches that is not easily filled. The gap spans the nerves that control muscle movement and muscle activation itself, or in other words, that exists between the central nervous system and the muscles controlled by it.

2.1.1 Speech production organs and the speech production system

Before we consider the gaps or omissions in the five-stage model further, we will first look at relevant components that are part of the five stages. This will help to provide background for a more detailed treatment of the specific theories of speech production in the following sections.

A brief look at speech organs is essential at this point, not only to understand what we might consider speech organs but also, crucially for speech initiation, to

understand the temporal properties of different parts of the speech production system. Particularly relevant is how the temporal organisation might work when the speech production system transitions from non-speech to speech. Knowing the limitations of the different parts of the speech system tells us which factors play a role in the behaviour of the system as a whole.

Let us first consider a very broad definition of what we mean by speech production organs. Since a review of a number of textbooks did not yield a clear definition of either speech production of organs or speech articulation, we use instead one which is based on a synthesis of implicit definitions available in various sources (Chiba and Kajiyama 1941, Fant 1960, Ladefoged 1995, Stevens 1998, Hewlett and Beck 2006, Gick et al. 2013). When implicit definitions were given they were usually in the form of a figure such as Figure 1 which opens the seminal work by Chiba and Kajiyama (1941). The figure presents what the caption calls 'vocal organs' – a midsagittal view of the vocal and nasal tracts with the larynx. In acoustic terms, it does cover the parts of anatomy that are essential for speech production. In aero-acoustic terms, the lungs and respiratory muscles are missing, and in terms of motor control and psycholinguistics, the neural structures – including the brain – are missing. Finally, one might wish to add feedback structures like proprioceptive nerves and hearing organs.

In general, a wide definition of speech production system should contain at least the relevant parts of the nervous system, the respiratory system, the laryngeal and supralaryngeal parts of the respiratory airway without forgetting any mechanisms that move the walls of the airway.

This means that in terms of the nervous system we need to consider the brain, the rest of the central nervous system, the peripheral nerves controlling the rest of the speech production system, as well as various sensory feedback systems – including all of the hearing organs – to be part of the speech production system. As for the respiratory system, the lungs, all of the respiratory muscles, and the trachea should be considered to be part of the speech production system. For the laryngeal system and supralaryngeal system, all of the laryngeal valve system, vocal tract (pharyngeal and oral parts) with connecting structures such

as the tongue and lips, and furthermore, the nasal tract and any connected sinuses (even if they only play a minor, passive part in sound transmission) should all be regarded as parts of the speech production system.

For the scope of the current thesis, a narrower focus is needed. For this reason we will review most of the speech production organs, but will omit some components from more detailed theoretical review and further components from the experimental focus of the thesis. Processes of the brain and central nervous system are of interest in the context of this thesis only in view of the timing information that is present in an executable speech plan and in view of how the anatomy of the neural system affects timing of motor control. For this reason, we will not discuss brain processes in detail. While breathing is certainly of interest in speech initiation, the focus of the experiments in this thesis is tongue articulation and so the respiratory system will only get attention when discussing the potential limitations it may impose on the timing of speech initiation.

Before we move on to discussing minimal reaction times of various parts of the speech production system, we will take a look at the nervous system and its anatomy.

2.1.2 Neural control structures and anatomy

As noted earlier, parts of the central nervous system (certain brain areas and parts of the spinal cord) are also part of the speech production system. They are after all the components that control speech production by planning and executing commands for articulator movement and processing signals from various feedback systems to make corrections and to accommodate the system to changing circumstances. In this thesis however, the focus is on speech initiation that takes place after planning and motor command generation are completed, so processes that happen before motor command execution are only briefly discussed.

As mentioned earlier, there seems to be a gap in theories and models, when we move from neural activity to physical activity: Models that provide detail

on neural processes either say nothing about the physical activity involved with speech or give only limited attention to it (Sternberg et al. 1988, Levelt et al. 1999, Guenther et al. 2006) while acoustic and articulatory models usually pass speech planning completely by (Fant 1960, Mermelstein 1973, Birkholz 2005) even if they do implement the control of muscle activation as a neural process (Lloyd et al. 2011). As we will see next, this is in part due to the very complex way the central nervous system controls articulation.

For speech to be possible, different speech organs need to be moved and their movements carefully coordinated relative to each other. To do so the central nervous system needs to trigger each of these events, but the length of the pathways from the brain and the spinal cord to the relevant muscles varies greatly (Lenneberg 1967). In particular, the recurrent nerve, which innervates *part* of the laryngeal system, is at least three times as long as the branch of trigeminal nerve that controls certain jaw muscles. Another part of the innervation of the larynx is provided by the superior laryngeal nerve. Like the recurrent nerve, it is a branch of the vagus nerve. However, unlike the recurrent nerve, the superior laryngeal nerve takes a more direct, shorter path to the larynx (Fuller et al. 2012), thus making the situation complex.

Not only does the length of nerves vary significantly, but signal transmission speed in the nerves varies between individual nerves based on their diameter (the thinner the nerve fibre the slower the signal) and Lenneberg (1967, referring to earlier studies by Krmpotić 1958; 1959) concludes that the innervation time differences for different speech articulators can be as long as 30 ms depending on the individual speaker's anatomy. Thus, in order to achieve the right temporal order of muscle activations at the ends of the various peripheral nerves, the central nervous system needs to send the triggering neural signals in a different temporal order to account for the differential transit times.

By comparison, articulation rate is in the range of 4-6 syllables per second (Lee and Doherty 2017) producing a speech segment roughly every 100 ms (assuming a CV syllable structure). Very fast speakers are able to reach verified rates as high as 10 syllables per second (Jannedy et al. 2010) producing a speech

segment every 50 ms. This means that at the fastest rates and depending on the individual speaker's anatomy, it may be necessary for the neural control signals to parts with longest and shortest innervation to be sent in with an offset that is close to segment length if not actually longer. Given that this offset varies from one speaker to the next, it is not surprising that building an explicit model for how the brain controls speech movements is not easy. Furthermore, it should be noted that the transmission estimates reported by Lenneberg (1967) are *not* based on direct measurements of transmission speed of nerve signals in the vagus nerve but rather are an extrapolation based on size statistics of the vagus nerve. As direct measurements of nerve signal delays lie outwith the scope of this thesis, in the current context we will have to rely on other measures such as the reaction times of speech related tasks that will be discussed in Section 2.1.4 below.

2.1.3 Phonation and respiration in speech

Phonation is achieved by a complex interaction between airflow physics, soft matter physics in the mucosal membrane and other structures of the vocal folds, and acoustic phenomena of air flowing out from the lungs through the glottis into the vocal tract above (Titze 1980; 2008). The end result of these interactions (all three subdomains are interconnected) is that in modal phonation the vocal folds, driven mainly by air flow, vibrate in a non-trivial manner, and in vibrating periodically close off the airflow, which gives rise to the glottal pulses, which are the acoustic result of phonation.

From the point of view of the respiratory system, phonation is achieved by using respiratory muscles to regulate the flow of air out of the lungs. Both active exhalation (driven by contracting muscles) and passive exhalation (driven by relaxing muscles) play a role in speech production. Further, in order to regulate airflow the inhalation muscles are also recruited to slow down the rate of exhalation. If the speaker adducts the vocal folds at the same time as exhaling air from the lungs, the vocal folds start to vibrate which results in phonation. Achieving this requires simultaneous control of respiratory muscles – to control

subglottal pressure – and laryngeal muscles – to fine tune the conditions so that vocal fold vibration can be sustained.

The relationship of respiration and speech is different from the relationship of most of the other functions of the speech organs and speech. This is because speakers still have to breathe while speaking for the purpose of keeping the oxygen/carbon dioxide exchange active. At the same time, respiratory activity is required for the production of speech sounds. So, while other functions – such as mastication and swallowing – can be, and in most cases need to be, suspended while speaking, respiration needs to be maintained – even though there can be momentary pauses. Many changes in respiration do not produce sound and can be implemented without interfering with turn signalling or other speech communication acts. Thus, respiration is adapted to rather than replaced by speaking. This makes the timing of respiratory activation different from the timing of articulatory activation, and also makes analysis of respiratory data different from analysing most speech production data.

Analysing the complexities of respiration and phonation in speech initiation is outwith the scope of the current thesis. While both voiced and voiceless sounds are used in the experiments reported in following the chapters, phonation and respiration will not be accounted for.

2.1.4 Temporal limits

As we will see, much of speech initiation research uses speeded trials or reaction time type tasks, and this thesis is no exception. When we seek to understand the timing structure of speech initiation, we are also seeking to understand what might potentially delay or stop the speech production process from starting. Because all parts of the speech production system may affect a given speech reaction time, it is important to know their minimal response latencies. Some of these are derived from reaction time tasks that are distinctly non-speech-like in nature while some come from speech tasks.

Section 2.2.1 will take a closer look at relevant speech reaction time paradigms, but in the meantime we should introduce two experimental

paradigms, that are used by studies cited in this section:

Minimal reaction time is a reaction time experiment paradigm where the reaction time consists of nothing but the time it takes for the participant to detect the 'go' signal, trigger the required action, and start moving.

Delayed naming is a speech reaction time experiment paradigm which is the speech equivalent of minimal reaction. The speaker is given the target word or utterance (usually in writing) and given time to prepare before the 'go' signal.

It should be noted that immediate naming – naming as soon as possible after the stimulus word appears – is not a minimal reaction time, because there is a task (reading and phonological/phonetic encoding) that has to be done before the response can be initiated. In minimal reaction time it is essential that the only requirements are detection of the 'go' signal and triggering of the response.

Table 2.1 at the end of this section gives a summary of the minimal reaction times of various speech-related gestures, that is, the fastest response speeds of some of the components of the speech production system. The table is organised following the airway from the lungs to the lips, and besides the reaction time, also lists the type of stimulus used to elicit that reaction time. Where possible, in the following review of studies of minimal reaction times, we will choose data obtained with an auditory stimulus, because that is the most common way of signalling to a participant to respond in the studies reviewed below. Section 2.2.1 relates the minimal reaction times to speech reaction times from various speech tasks.

We will first take a look at some relevant properties of both neural and physiological subsystems that are known to be relevant for speech from neurological and physiological studies. We will then look at articulator specific overall response latencies and finally, in the summary, discuss the implications of what we know about the timing of speech initiation.

Maximum movement speeds of different articulators would have to be carefully related to the speakers anatomical measurements to provide additional timing information, and are thus beyond the scope and methodological means

of this thesis. Instead, we will rely on the timing information available in the form of reaction times.

Relevant properties of subsystems

This section takes a brief look at the time taken by necessary processing stages from acoustic stimulus onset to articulatory onset of the speech response and for the acoustic onset of speech to be registered. The section then discusses the temporal properties of muscle fibre activation.

Chiu and Gick (2014) calculated an estimate of the minimum time required for a speech response when triggered by an auditory stimulus. Their study used the STARTLE paradigm which will be discussed in Section 2.2.1. The following is an adaptation of their calculations for the purposes of this thesis: It takes 10-25 ms from the onset of acoustic stimulus for activation to reach the auditory cortex (Schroeder and Foxe 2002). From there, different cortical processes leading to activation of primary motor cortex take a total of 7-14 ms (Stockard et al. 1977, Guenther et al. 2006, Carlsen et al. 2012). Finally, the orofacial muscle Electromyography (EMG) response to trans-cranial magnetic stimulation is added. It is 11-12 ms according to Meyer et al. (1994) and followed by a motor time (EMG response to movement) of 30 ms. These estimates will be used for later chapters to provide a conservative lower bound used in removing outliers from experimental data. For now, it is sufficient to note that the sum of all of these delays is 58-81 ms.

For completeness, it should be noted that sound does not travel instantaneously. The time that it takes for a sound signal to propagate from the glottis to the lips is negligible. With a commonly assumed speed of sound of 350 m/s and a long (conservative estimate) vocal tract of 20 cm length, the lag would be $t_{lag} = .2/(350) \approx .57$ milliseconds. However, for a microphone placed at the distance of one meter from the lips, this becomes $t_{lag} = 1.2/(350) \approx 3.4$ milliseconds, which no longer is necessarily negligible. A microphone should thus be placed no further than a meter away from the speakers lips, which is common practice in speech laboratory recordings.

Articulatory muscles are of the skeletal type as opposed to the smooth muscles of blood vessels and intestinal walls, and the cardiac muscles, which constitute their own type (Feher 2012). The smallest separately innervated part of a skeletal muscle is a muscle fibre, which is controlled by one or more axons of a motor neuron. Importantly from the point of view of initiating movement, a single neural activation signal to a muscle fibre produces a twitch, which consist of a latent period, contraction, and relaxation. If the activation signal is repeated before the fibre has time to relax, the fibre will produce a stronger contraction as a sum of single twitches. The latent period is up to 10 ms long, and the overall duration of the twitch is usually in the range between 10 and 100 ms. Both depend on the physiological characteristics of the individual muscle fibres.

Minimal reaction times of articulators

Draper et al. (1960) report minimal breath reaction times from one participant in the form of EMG activation of internal intercostal muscles in response to a delayed naming trial with the target word 'ma'. Unfortunately, they do not provide deviation values. They only state that the shortest internal intercostal EMG reaction time was 140 ms and the longest was 320 ms. Draper et al. (1960) also report the delay from internal intercostal activation to acoustic onset as a relatively constant interval of 48 ms (sd = 8 ms, n = 9).

Izdebski and Shipp (1978) measured minimal reaction time for voicing onset. They accounted for the effect of stimulus type (auditory and somatosensory – we use the results for auditory stimulus here as mentioned above), vocal fold starting position (abducted or adducted), amount of air held in the lungs (1/4, 2/4, or 3/4 of vital capacity), and possible gender effects. The effect of amount of air held in the lungs was statistically non-significant, but the rest of the variables did have significant effects. The values displayed in Table 2.1 are those for the male speakers with vocal folds abducted before phonation onset. The finger lift reaction time provided in the table as the baseline comes from this same study, but since there were no gender differences, we use the pooled data of both female and male participants.

Table 2.1: Means and standard deviations (sd) of minimal reaction times of various speech-related gestures in milliseconds. See text for descriptions of the experiments used. ('na' stands for not available, the value was not listed in the report; ms for milliseconds, RT for reaction time, 'sd' for standard deviation.)
*Results from two participants, voice key based mean reaction times from same participants are 168 and 170, which points to the labial opening being recorded later than the onset of acoustic speech.

Gesture	Stimulus	Mean RT (sd) ms	n	Reference
Finger lift (baseline)	Auditory	149 (29)	30	Pooled data from Izdebski and Shipp (1978)
Internal intercostals	Auditory	140 (na)	n	Draper et al. (1960)
Voicing onset	Auditory	170 (32)	15	Izdebski and Shipp (1978)
Velar closure	Auditory	206 (62)	297	Dalston and Keefe (1988)
Tongue tip press	Auditory / 10 dB	209 (94)	n	Siegenthaler and Hochberg (1965)
Tongue tip press	Auditory / 50 dB	137 (56)	n	Siegenthaler and Hochberg (1965)
Tongue tip press	Auditory / 79 dB	129 (55)	n	Siegenthaler and Hochberg (1965)
Tongue tip press	Tactile	123 (47)	n	Siegenthaler and Hochberg (1965)
Labial closure	Auditory	203 (95)	290	Dalston and Keefe (1988)
Labial opening	Visual	188 & 176 (na)*	130 & 130	Cattell (1886)

Mean reaction times of the tongue to tactile and auditory stimuli were obtained by Siegenthaler and Hochberg (1965) with a purpose built biteblock system. The device could measure tongue tip presses, where the tongue is used to press a button built into the biteblock, and also functions as a a stimulus source by providing a tactile go-signal to the participant by vibrating.

Dalston and Keefe (1988) measured velopharyngeal and labial reaction times with a photodetector system and used auditory stimuli in their experiments. For velopharyngeal measurements a photodetector was inserted into the participant's nasal tract, and for labial measurements it was attached to the participant's lower lip. In both cases the light source was inserted through the nasal tract into the participant's pharynx.

Summary

Only tentative conclusions can be drawn from the data presented above because different studies used different participants, and we do not have access to the raw data to carry out statistical comparisons. In some cases there are not even error bounds available. However, certain trends do become apparent with careful consideration.

Looking at the data summarised in Table 2.1, we can see that, while all of these estimates are clearly slower than the lower bound estimate of 58-81 ms provided by Chiu and Gick (2014), some articulators seem to be considerably slower to respond than others. Humans respond fast with their tongue to both auditory and tactile stimuli (as long as the auditory stimulus is loud enough). Intercostal muscles responses are almost as fast as tongue responses which are both faster than finger lift response which is often used as a baseline in reaction time studies – though equal within the error bounds available.

That voicing onset exhibits a slower response time than respiratory muscles and tongue movement is hardly surprising, because exhalation flow and careful control of the position of the vocal folds are necessary for phonation. A possible additional explanation is the potentially long delay introduced by the recurrent branch of the vagus nerve innervating part of the larynx. Given the careful way

in which Izdebski and Shipp (1978) work through parameters that might affect the speed of a phonation response, the results from this study are among the most reliable and generalisable reviewed here.

In evolutionary terms the main function of the larynx is to prevent foreign objects and matter from entering the trachea which requires fast reflexive responses (closing the glottis or triggering a cough response), but which do not have to exhibit the kind of fine control that is necessary to produce phonation. Thus, it is natural to find relatively slow response times associated with a structure that has to have fast responses associated with its primary function.

Labial and velar reaction times being slower than any of the other groups might be considered surprising. However, it should be noted that in the tactile tongue reaction time trial the tongue would have been in contact with the biteblock, and all that was needed for the clock to be stopped was application of pressure against the same biteblock. Since the auditory reaction time from the same study is very similar, we can assume that the tongue was in the same position and these results can therefore be considered reasonably reliable. What we maybe see in the longer labial and velar reaction times, may in fact be the necessity to move the articulators further to stop the clock by closing the open velar port, or opening or closing the lips.

To summarise, we can with fair confidence assume that the tongue will be the fastest responding articulator in the vocal tract, or at least among the fastest, with the above caveats about task differences. We can also assume that the response time of respiratory muscles and either consequently or independently of them also the onset of phonation, will be what actually holds up the onset of acoustic speech. The role of lips and the velar opening remain unclear in the light of this data.

2.1.5 Feedback mechanisms in speech

Several feedback mechanisms or feedback loops are utilised in speech production. There is proprioceptive feedback about the position and motion of the muscles employed in speech production, haptic (or tactile) feedback from dif-

ferent parts of the articulators touching each other, aero-tactile feedback from airflow in the vocal tract and across the lips, and audio feedback from the sound produced by the speaker being heard by themselves (Gick et al. 2013). The first three types of feedback can be collected under the concept of somatosensory feedback. Speakers use these feedback mechanisms to ensure that articulators are behaving expectedly – matching articulatory targets closely enough – as speech is produced and that the auditory output matches the expected auditory targets, and implement corrections based on the information gained from them.

By artificially interfering with somatosensory feedback, researchers have found that speakers respond to somatosensory perturbations relatively quickly, with compensatory response times of 22-75 ms reported by Abbs and Gracco (1983) (summarised by Guenther 2016) and similar response onset times reported also by Gomi et al. (2002). In contrast, auditory perturbation studies have found that speakers start compensating for artificial auditory perturbations 100-150 ms after the onset of the perturbation (Purcell and Munhall 2006 summarised by Guenther 2016). Auditory studies have also found that speakers show larger responses, if the perturbation has the potential to affect the linguistic content of the utterance (Mitsuya et al. 2011, Niziolek and Guenther 2013 summarised by Guenther 2016).

From the point of view of speech initiation, feedback is certainly relevant, because it may trigger corrections and hesitations. Considering the above data, we would expect these to occur mostly through the proprioceptive loop, but possibly also through the auditory loop for longer onset sounds – Rastle et al. (2005) report onset consonant durations as long as 160 ms – and certainly during the whole of the onset syllable. Presence of corrections and hesitations would confound the goal of characterising the baseline speech initiation behaviour of speakers. For this reason, this thesis concentrates on unperturbed – artificially or otherwise, non-pathological utterances and excludes hesitations from most of the analysis. Thus, we can conclude, that while understanding feedback mechanisms is important in studying speech initiation in general, in this work they are of peripheral interest.

2.1.6 Articulation and acoustics

In this section we will first look at what is known about speech production acoustics and what conclusions can be drawn about speech initiation based on that knowledge. After that we will review two dominant articulatory models – ArtiSynth and VTLab – before discussing theories that span both articulation and acoustics in sections on coarticulation and on Articulatory phonology. The final section, before we move to models of neural control, reviews the speech ready position and related concepts.

Acoustics: Source-filter theory

The source-filter theory or model is relevant to this thesis because it gives a qualitative prediction on the relative timing of movement initiation and onset of acoustic output. The source-filter theory is an acoustic theory, which states that speech production is a process that has two essentially separable (that is, non-interacting) parts: the voice source and the vocal tract filter. Hence, the name source-filter model. It is a relatively old theory that was first formulated for vowels by Chiba and Kajiyama (1941), who described vowel production in great detail. They provided a description of the relevant anatomy and articulation, a very advanced acoustic model of the vocal tract, and an account of vowel perception. Their anatomical work was based on X-rays and plaster casts of Japanese speakers.

A more widely know treatment of the source-filter model is provided by the acoustic theory of speech production by Fant (1960), who based his work on data from an X-ray study of Russian. He expanded the work of Chiba and Kajiyama to include also consonant production by including the possibility of having a noise source present in the vocal tract. The most complete (and easily available) textbook treatment of the theory is by Stevens (1998).

In more precise terms, according to the source-filter theory the speech signal can be described as the product of three independent parts: the sound source, the vocal tract filter and the lip and/or nose radiation load. This can

be expressed as the Equation 2.1 (Rossing et al. (2002)) which is represented graphically and expanded slightly in Figure 2.2.

$$\text{Speech sound} = \text{Source} \times \text{Vocal tract filter} \times \text{Lip/Nose radiation} \tag{2.1}$$

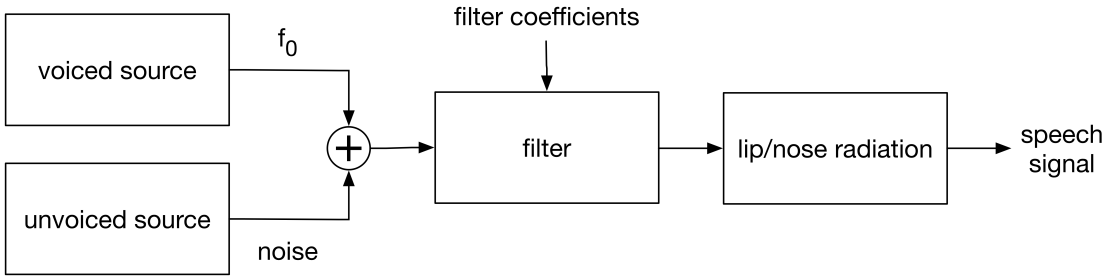

Figure 2.2: Components of the source-filter model of speech production. Please, refer to the text for explanation of the components. (Schematic drawn by the author based on Lemmetty (1999).)

The first part of the speech production model the source-filter theory is the source. As can be seen in Figure 2.2 the source can be divided into the voiced source, which refers to voicing produced by the laryngeal mechanism providing the fundamental frequency (f_0) for speech, and unvoiced source, which refers to the many possible sources of frication noise and transient noises within the vocal tract. At any given time during speech production one or both of these mechanisms may or may not be active. The sound is then filtered by the vocal tract's filter function, which is treated as the passive acoustic transfer function of the vocal tract, and by lip and/or nose radiation functions. The combination of all these components produces the acoustic speech signal that is transmitted to the surrounding environment.

If we consider the predictions of the source filter model from the point of view of speech initiation, we can see that there are three necessary conditions for the model to produce speech. First, there needs to be at least one active sound source in the system. This means that there are potential implications for breathing, larynx state, velar port, tongue, and lips depending on the sounds

being produced, and which of these structures the sounds require to act as active sound sources. Second, the filter needs to also be in a suitable position for speech production, which has implications for the oral tract, velar port, and the lips. Third, (for most sounds) the lips and/or the velar port need to be opened for any sounds to actually be transmitted to the outside world.

This means, that if the speech production system is not by coincidence in the right configuration to produce the first sound of the utterance, there is necessarily a lag between speech production being initiated and acoustic onset. In other words, according to the source-filter theory, when moving from non-speech to speech, we should observe a period of silent articulation before the acoustic onset of speech. This is leads to the timeline first presented in Chapter 1 in Figure 1.2.

A trivial case is provided by unvoiced plosives. They are sounds that need the vocal tract to be closed silently before the opening burst produces sound. It should, however, be emphasised that the source filter theory predicts that in almost every case of speech initiation regardless of what the phonetic content of the utterance is, there will be silent articulation before acoustic speech begins. As we will see in Section 2.2.2, the existence of a silent articulation phase is also verified by empirical data.

New acoustic models of the vocal tract have been developed in recent years (Dedouch et al. 2002, Hannukainen et al. 2007, Palo 2011) and in modern understanding the voice source and the vocal tract filter are no longer considered separable (Titze 2008). Yet, the basic prediction of the acoustic models stays the same: a speaker must move before producing sound.

Articulatory synthesis models

In this section we will consider two different articulatory synthesisers: ArtiSynth and VocalTractLab. ArtiSynth is a biomechanical speech articulation model or more correctly a biomechanical modelling platform (Lloyd et al. 2011). It is built to be modular. A numerical soft matter physics simulation kit makes up the low level modelling layer. Artisynth can be adapted to many purposes by

implementing anatomical components to model the structures that are being studied.

Artisynth's default model is a Three Dimensional (3D) model of the vocal apparatus which contains both volumetric structures such as the tongue and the mandible, but also point-wise muscles. An important thing to note is that the tongue is modelled realistically as a muscular hydrostat. This means that it is a volume preserving component of the model that is acted on by muscle forces.

The ArtiSynth model includes both tongue internal and external muscles acting on the tongue, and controls their movement based on the equilibrium point hypothesis (Vogt et al. 2006, Feldman 1986). However, the model does not currently have a higher level control component that would specify motor plan based on phonetic or similar input, but instead movement is synthesised by directly defining individual muscle activations strengths as functions of time. Neither is there an acoustic modelling component to ArtiSynth, and hence, it can not produce acoustic output.

In contrast, VocalTractLab (Birkholz 2005) is a 3D expansion of the classic 2D Mermelstein articulatory synthesizer (Mermelstein 1973) and able to produce sound via a numerical aero-acoustic simulation model, which is a modern descendant of the model originally proposed by Kelly and Lochbaum (1962). The movement of VocalTractLab is controlled by specifying vocal tract parameters that are an extension of the parameters of the original Mermelstein model. They are geometric in nature rather than biomechanical.

The strength of these models lies in being able to simulate the mechanics of speech production and provide predictions on how different control strategies would be realised by a physical speaker (Harandi et al. 2014) and providing powerful demonstration tools for learning about articulatory phonetics. In the case of ArtiSynth's level of physiological and physical fidelity, there is potential for the analysis-by-synthesis approach to provide reasonable estimates about how a copied speaker moved their muscles to achieve the observed movement sequence. In contrast, VocalTractLab's strength is that it can provide similar estimates of what a given vocal tract configuration will sound like. However,

the way that these models have been constructed illustrates the divide between models of neural processes and models of physical processes – albeit that ArtiSynth does contain a low level neural control model, but neither of these models includes a model of utterance planning on any level other than manual specification.

Both models use a rest position as their origin of articulation, but they differ in how they define it. VocalTractLab uses the schwa ([ə]) as the rest position and therefore as the starting point of articulation (Birkholz 2005). From the point of view of speech initiation, this is a problematic assumption since without actually knowing that this is what speakers do, we can not rely on the model to provide accurate simulations of speech initiation. On the other hand, ArtiSynth defines the rest posture as a position with no muscle activations (Stavness et al. 2011), which is physically reasonable, but has an implicit assumption that when 'resting' speakers do not engage in any other activities such as breathing or that breathing is facilitated optimally by relaxing all articulators involved in speech. Again, without actual data on the nature of the rest position, we can not say whether the assumption is accurate. This thesis provides behavioural results on the timing of the transition from rest to speech, but the topic of what rest positions, ready positions and related concepts clearly merits further study in the future.

Coarticulation

Speech sounds are seldom produced in independent isolation and when they are produced with other speech sounds they affect the way each of them is realised. This means that when a speaker initiates speech, it is not enough to know how the first sound will be produced, but instead we have to consider how it will be produced in the phonetic context that it appears in.

As a general example, in a front vowel context, the closure of [k] is usually further front than in a back vowel context. Coarticulation is the name usually given to this phenomenon (Kühnert and Nolan 1999). Coarticulatory effects are language specific (Manuel 1999) and also depend on factors such as the

speaker's age (Fougeron et al. 2018).

In time, coarticulation can be understood to consist of two separate mechanisms: look-ahead coarticulation and carry-over coarticulation (Farnetani and Recasens 1999). The latter refers to the effect of the articulation of a speech sound on following sounds. In speech initiation, this is of little concern as there are no preceding sounds before the first one.

In look-ahead coarticulation, some articulatory gestures needed to realise a given speech sound begin during the production of previous speech sound or sounds (Farnetani and Recasens 1999). For example, when producing the English utterance [kɔːtʰ], a speaker will most likely begin the lip rounding gesture of [ɔ] before the articulatory release and, thus, before the acoustic onset of [k]. In a speech initiation context, this means that the first structure to move is not necessarily one that is directly associated with the articulatory target of the first phoneme. Consequently, when analysing timing data, the phonetic identity of sounds following the onset sound should also be accounted for.

Articulatory Phonology

Articulatory Phonology is an important theoretical and modelling framework for temporal organisation of speech (Browman and Goldstein 1990; 1992). Articulatory Phonology aims to bridge the gap from phonological encoding – that is, from the phonological output of the mental lexicon, to actual articulator movement. It proposes that stored phonological units are actually articulatory in nature. This view makes it unnecessary to map the phonological representation of the utterance first to a phonetic representation and then to actual articulator movement targets. Instead, the phonological representation directly specifies the articulatory targets and time windows when they are active. The articulatory mechanism (both neural and physiological elements) is responsible for assigning trajectories to each target and for blending the gestures together to produce utterances.

A key concept that has bearing on timing of pre-speech in the definition of syllable timing in Articulatory Phonology is the C-center or consonant center

(Browman and Goldstein 1988). It describes the way that gestures of individual consonants and consonant clusters are timed in relation to the vowel nucleus of a syllable. Since Experiments 2 and 3 of thesis will use single consonant onsets, this is particularly interesting in the case where the consonant is a single the onset of an utterance.

For an individual consonant, the C-center is defined as the center of the steady state of the consonant's articulatory gesture. The steady state is equated with the plateau in the consonant gesture where the articulator that forms the constriction has reached maximal displacement. For a consonant cluster, it is defined as the mean of the individual consonant gestures' center points. According to Browman and Goldstein (1988) the distance between the C-center and the attainment of the final consonant's closure in a /C(C)(C)VC/ word is the least variant of the various timing relationships. This means that, if the initial consonant or consonant cluster grows longer, it will both cut into the length of the vowel and start earlier.

Since the C-center is effectively the center of the maximum constriction of the consonant, it will also correspond roughly to the center of the acoustic segment for most singleton consonants. The obvious exception are plosives, because the release burst happens in the release period of the gesture after the constriction plateau has been passed, and aspiration happens only after the release burst.

In terms of speech initiation, the prediction depends on whether the length of the onset period is constant, or if its length is affected by the length of the consonant constriction plateau. This in turn is a complex question that is left open at least by Browman and Goldstein (1988) (see endnote 3). This is hardly surprising since the evidence for C-centers comes from analysing the timing of target words embedded in a carrier sentence where the offset of preceding phonemes will blend with the onset of following ones. We will revisit this prediction when discussing speech initiation results in Section 2.2.3.

Also relevant to speech initiation, the existence of a speech ready position has been hypothesised in the context of Articulatory Phonology by at least

Saltzman and Munhall (1989) and Simko (2009). In Articulatory Phonology it is an articulatory state from which all speech movements and gestures begin and to which the articulators return after an utterance, if further speech is anticipated. Since this hypothesis is not unique to Articulatory Phonology, it will be discussed in more detail in the next section.

Speech ready position

There is a long held view that speech movements are based on a basis of articulation or an articulatory setting – a neutral position which all actual speech movements are built on. It is specific to the speech mode (Wilson 2006) and thus separate from being at absolute rest for respiration or activities such as mastication and swallowing. Historical surveys of the concept are provided by Kelz (1971), Laver (1978), and Jenner (2001). The concept dates back to at least the 17th century: Wallis (1653, cited by Van Buuren (1995)) talks about the differences between the ways English and German are produced – not just in terms of a different sound inventory but in terms of a different overall articulatory setting.

While there are many interpretations of the basis of articulation or what might be considered related concepts, the most relevant concept in the current context is the speech ready position. The speech ready position or a neutral articulatory position which is optimal for producing a given language. It is an appealing concept in speech modelling, because it gives parts of the articulatory system a default position to move to when no articulatory targets are defined for them. It is part of the Task Dynamical implementation of Articulatory Phonology (Saltzman and Munhall 1989, Browman and Goldstein 1992) and its descendant, the Embodied Task Dynamics model (Simko 2009), as the neutral position of the coarticulatory model.

Yet until fairly recently, the ready position has remained only a theoretical concept. This is because measuring it has proved to be quite problematic (Heffner 1950, O'Connor 1973, Collins and Mees 1995). However, in his study of both monolingual and bilingual English and French speakers, Wilson (2006)

managed to analyse the static qualities of the inter-speech posture which he identifies as a close cognate of the speech ready position. Wilson recorded the articulation of speakers, who were waiting for more text to read, with both an optical point tracker device and tongue ultrasound. He found that in his task speakers employed a steady position that lasted over 300 ms in about 50 % of the tokens with the rest of the tokens containing movement throughout. Unfortunately, the text only gives the relative differences in the inter-speech postures and does not provide an explicit description of the articulatory positions employed by the speakers.

The pause length that is required for a speaker to use the speech ready position was studied by Schaeffler et al. (2008) who analysed Electropalatography (EPG) data from three speakers. Like Wilson Schaeffler et al. do not give a characterisation of what the actual articulatory postures used by the speakers were but only give a definition of an inter-speech posture as a time when the EPG data did not show "any notable change in the overall contact pattern" Schaeffler et al. (2008). They tested the occurrence of the ready position with pause lengths of 2, 5, and 8 seconds and found that the three participants were most likely to use a speech ready position when the pause was 5 seconds. Wilson's experiments had a constant 2 s wait before the next token, which may have resulted in the participants learning the length of the wait and adapting to that. It certainly resulted in suboptimal occurrence of the speech ready position as shown by Schaeffler et al. (2008).

From the point of view of pre-speech or speech initiation – that is, moving from a mode other than speech to speech mode – the speech ready position is less important. Assuming that a neutral position would appear at the beginning of speech initiation implicitly assumes a system, which is never used for anything but speech. This makes sense, if the system in question is an articulatory speech synthesiser, but less so if we are talking of an actual human being. On the one hand, if a speaker does know what they are going to say, only experimental settings – or some other task taking priority over speaking – will prevent them from preparing their utterance articulatorily. Only in the case that the speaker

does not know what they are going to say, but they have pressure to start as soon as they can when they do know – like Wilson's experiment – is there time left for a neutral position such as 'speech ready position' to be an intermediate *and held* target before utterance articulation begins. However, since the eliciting of the speech ready position proved so difficult before experimenters started using the inter-speech pauses as the experimental paradigm, it would seem unlikely that we would be able to reliably observe the speech ready position in a task where the participant is moving from another mode like resting to speech.

2.1.7 Models of neural control

In this section we are now going to take a closer look at three representative models that deal mainly with the first three stages of speech production as set out in Figure 2.1. They all deal with aspects of utterance planning. They are introduced to exemplify the divide between models of neural and physical processes. The first model also provides a prediction on how utterance complexity affects the acoustic response latency.

The first model is an influential model of rapid sequential linguistic production that provides predictions about the timing of utterances of differing complexity. The remaining two models expand the planning and motor command stages with finer structure: For planning, deciding what needs to be communicated, which utterance to use to achieve that, and when and how to produce the selected generated utterance. For motor commands, integration of linguistic and para-linguistic content, as well as accommodating for contextual variables such as background noise or restricted articulator movements.

Model of rapid speech

Sternberg et al. (1978) and more recently Sternberg et al. (1988) proposed a model of motor control in *rapid* speech. The model is based on delayed naming experiments or minimal speech reaction time tasks. Production targets were words or word lists, and both number of syllables and number of words in the

lists were systematically mapped in the different experiments (Sternberg et al. 1978). The model is illustrated in Figure 2.3. We see that the model assumes that there exists a motor program for the whole prepared utterance and that this program is made up of sequentially linked units.

First of the two main axioms of the model is, that speech control alternates between speech program unit selection and program execution. The second axiom can be stated as that if the system is pushed to the limit by requiring the speaker to be as fast as possible, the limiting factor on speech rate will be the time that it takes to go through this cycle meaning that each action unit of speech will – at the limit – be the length of the cycle. In the model, an action unit is assumed to be the prosodic phrase or stress group. The model predicts longer action units (ones containing greater number of syllables) will have a longer response latency in comparison to simpler ones meaning potentially a longer pre-speech duration.

The model attributes speech and articulation rates in rapid speech totally

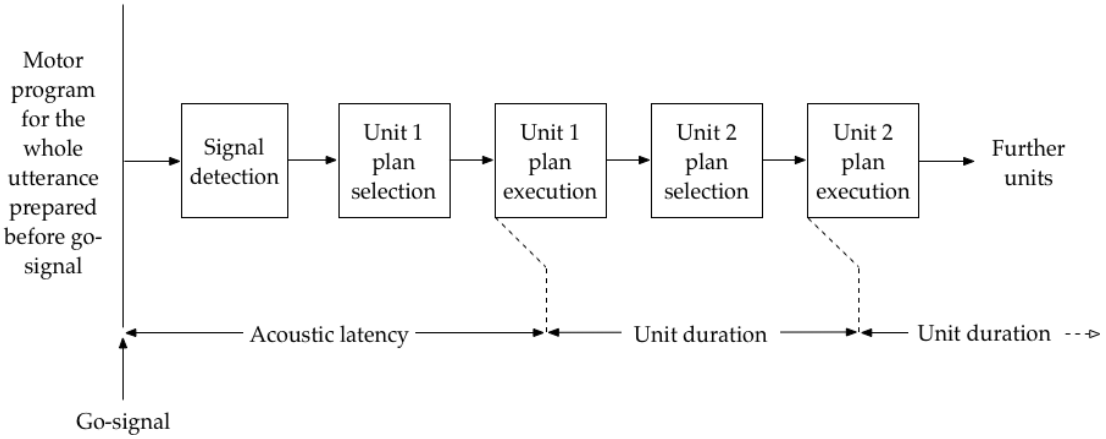

Figure 2.3: The model of rapid speech production by Sternberg et al. (1988). The speaker is first given time to prepare the whole target sequence after which they wait for the signal to initiate production. This is followed by a signal detection latency after which the first speech unit is selected from the motor program, and executed. The model then moves to the next unit and repeats the process until the whole target sequence has been produced. (Graphic by the author based on Sternberg et al. 1988, p. 184).

to neural delays of motor program retrieval and execution. It thus completely disregards the possibility that the limiting factor might be either the inertia of the articulators or a physical constraint on how fast intelligible speech can be produced by the articulators.

Crucially from the point of view of pre-speech, the model does not specify what happens between executing commands and output. However, the model does (or its authors do) implicitly acknowledge that there is a delay between neural execution of a speech command and acoustic output. This is shown in Figure 2.3 as slanted lines from beginning of a unit command to beginning of acoustic output.

From a phonetic point of view the model is based on speech experiments which are not ideal. The required onset and offset times were determined by a voice key, which is an automated sound intensity based device. Voice keys have been shown to be sensitive to changes in the phonetic identity of the target utterance onset. Since the phonetic onsets of target sequences varied within the experiments – some of the experiments have a combination of vowel onsets and different consonant and consonant cluster onsets, the results may not be totally reliable. Furthermore, as we will see in Section 2.2.2, the onset phoneme of an utterance causes an actual variation in the acoustic onset latency in a minimal reaction time or delayed naming task. Voice keys will also be discussed in Section 2.2.2.

WEAVER

WEAVER is model of utterance planning developed by Levelt et al. (1999) and later expanded by Indefrey and Levelt (2004). A schematic of the earlier version of the model is shown in Figure 2.4. The earlier version of WEAVER is based on reaction time studies and studies of speech errors (Levelt et al. 1999) and the update on a comprehensive meta-analysis of both word production and perception studies including a large number of brain activation studies (Indefrey and Levelt 2004).

Even the older, simpler version illustrated in Figure 2.4, is detailed in break-

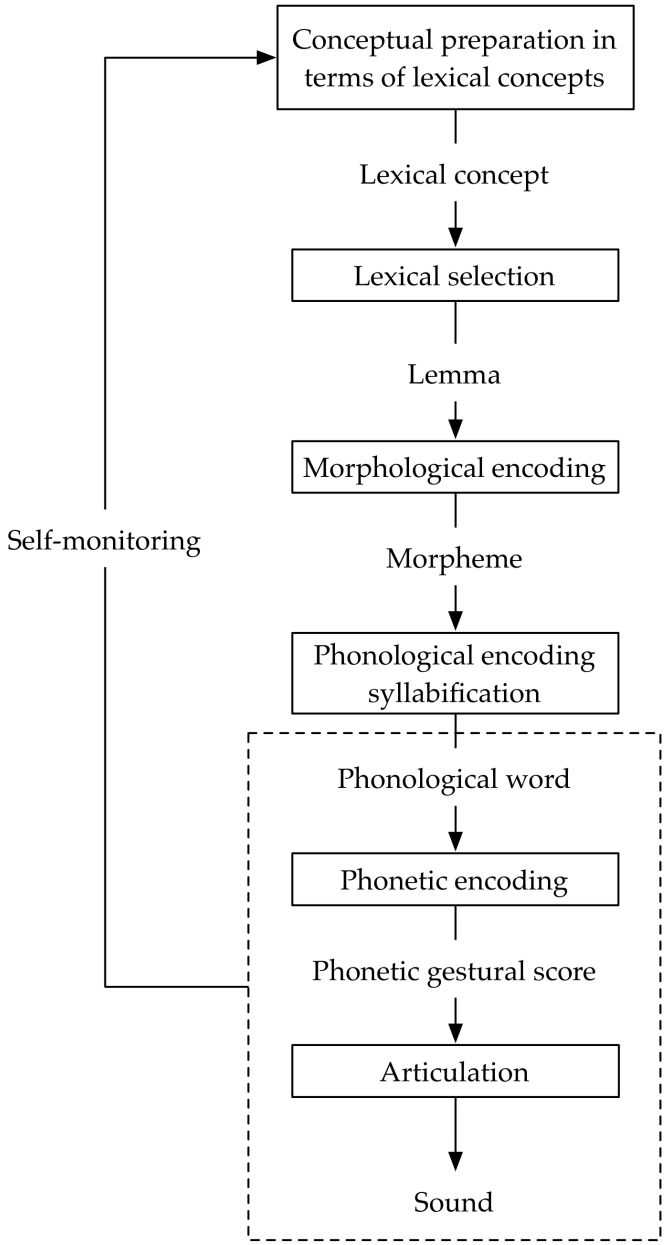

Figure 2.4: The stages of lexical access and speech production according to the theory by Levelt et al. (1999) also known as the WEAVER model. (Levelt et al. 1999, Graphic by the author based on)

ing down the concept-to-speech arch into individual processing stages. The expansion reported by Indefrey and Levelt (2004) links the processing stages to different speech tasks and relates the model components to brain processes. It also adds stages needed to explain word reading, word listening, and picture naming, along with interconnections and processing delays of the different processing components to explain the data observed in the studies it is based on.

From the point of view of the division to neural and physical activity (Figure 2.1), WEAVER is well accomplished on the neural side but extremely underspecified on the physical side, which is represented only by the 'Articulation' component in both versions of the model. Even though the schematic includes the phonetic gestural score generated by phonetic encoding, this is not a gestural score that would include gestural timing information. Neither is there any specification of how the timing information would be produced by the articulation process. As it is, we must conclude that the WEAVER model will be extremely interesting in predicting the time course of more complex speech tasks and their variations, but in the current context of looking at timing of pre-speech articulation, it does not provide any predictions of the relationship of articulatory timing to acoustic timing.

DIVA and Godiva

The DIVA (Directions Into Velocities of Articulators) and its extension GODIVA (Gradient Order DIVA) are speech production models whose focus is on computational modelling of cognitive processes and learning (Guenther et al. 2006, Bohland et al. 2010, Guenther 2016). They model the neural processes responsible for controlling speech production with artificial neural networks designed to imitate the processing delays and functioning of both verified and hypothesised neural structures in the brain and central nervous system. Like WEAVER, GODIVA and DIVA rely on extensive meta-analysis of speech production and perception studies in defining the model.

There are however two key differences to the WEAVER model. First,

WEAVER originally started as a more abstract model (Levelt et al. 1999) and only later started relating the processing components more closely to neural structures (Indefrey and Levelt 2004). In contrast, already in DIVA we have a direct relation of model components corresponding to neural substrates. Second – and more importantly in the present context, DIVA, and thus GODIVA, includes a model of articulation and acoustics for actual acoustic output.

The model of vocal tract and its acoustics is based on an articulatory synthesiser by Maeda (1982; 1990). The articulatory model is a parametrised model rather than a physiological one. The vocal tract movement parameters are derived from articulatory data by factor analysis. The acoustic synthesis is performed by an implementation of the source-filter theory with a numerical framework described by Maeda (1982) and derived from the acoustic model proposed by Kelly and Lochbaum (1962).

GODIVA is quite likely the most detailed currently available model of the neural structures that control speech production and of motor learning in speech production. Since its implementation is based on copying actual neural structures, it gives interesting information on how the processing structures might be organised in the brain and central nervous system. At the same time, the wealth of structural detail that GODIVA provides makes it difficult to use it as a predictive model. Unlike simpler – for example, purely rule based – models GODIVA is less explicit with no outright defined time domain behaviour. To provide concrete predictions of articulatory and acoustic timing the model needs to be adapted to each speaker before running simulated production trials (Nieto-Castanon et al. 2005).

For the purposes of this thesis, it is more efficient to record data from an actual speaker than adapt GODIVA to several speakers as well as record data from them. However, in future a model like GODIVA – possibly in combination with a more realistic articulatory synthesiser such as Artisynth – can potentially be used in bridging the divide between models of neural and physical processes.

2.1.8 Summary

From the literature reviewed in this section, we now have estimates of how long it at the very least takes for speech to be initiated from the analysis by Chiu and Gick (2014). Based on the estimates of how long it takes to initiate movement with different articulators when they are used independent of speech, we expect that from a physical point of view the initiation of phonation might be the dominating delay in speech initiation.

We have also reviewed acoustic, articulatory and neural control models and theories. While a lot of the theories are ambiguous in respect to pre-speech and speech initiation, the source-filter theory predicts that we will see speakers move before the onset of acoustic speech and theory of C-centers in the Articulatory Phonology framework, while non-specific about speech initiation, does predict that the centers of a word initial consonant constrictions of different durations will remain equidistant from the end of the first syllable with all other things remaining equal.

While the review of acoustic theories, articulation and neural control models above provide a starting point for understanding speech production as a process, there are non-trivial problems in understanding the speech production process as a whole that are left open by them. Specifically from the point of view of this thesis – that is, the point of view of speech initiation studies – the temporal organisation of pre-speech is still left vague. As Sternberg et al. (1988) (page 177) say: "... by forcing performance to its limiting speed we are more likely to discover its fundamental constraints." So let us next take a look at speech initiation studies and speech reaction time research.

2.2 Speech initiation research

Pre-speech articulation is movement that prepares the vocal tract for producing acoustic speech. Pre-speech can be divided into non-linguistic and linguistic preparation, that is, preparatory movements for getting ready to speak and silent speech movements that lead to acoustic speech being produced. Figure 2.5 shows the preliminary conceptual timeline of pre-speech, which was first introduced in Chapter 1. This timeline will be refined in this chapter based on earlier research and theories, and later further refined in the Discussion (Chapter 7) based on the results of this thesis.

Preparatory movements might include actions such as inhalation, licking of lips, and opening the mouth. In contrast, silent speech movements associated with the first segments of speech production are utterance-initial movements, that is, already part of the speech articulation. Silent speech movements will reflect prosodic structure and paralinguistic intentions as well as lexical content, and can be expected to be language specific, that is, truly linguistic.

Mental chronometry is the use of reaction time to infer properties of cognitive processes. It has been actively used by psychologists since the 19th century (Jensen 2002). However, the oldest known definition of the concept of reaction

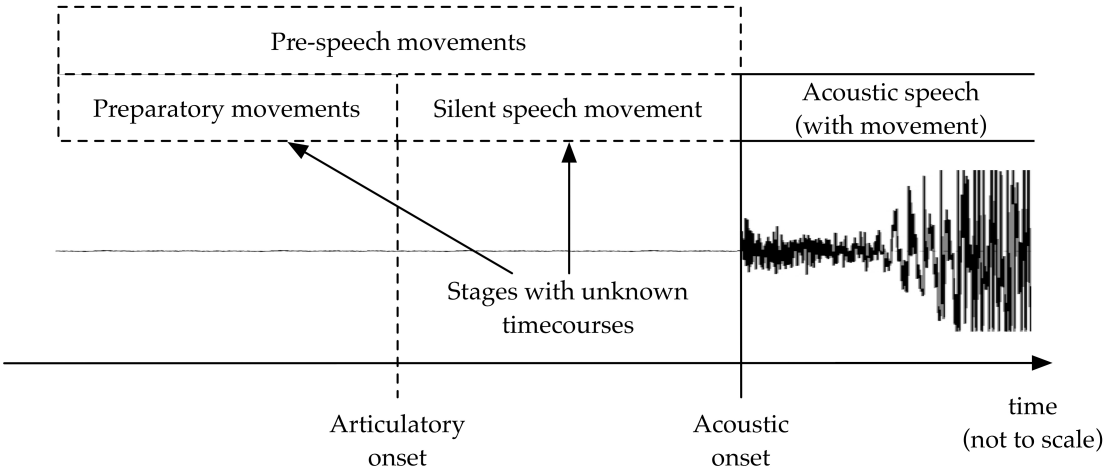

Figure 2.5: Conceptual timeline of pre-speech articulation and onset of acoustic speech.

time comes from "Optics" by Abu Ali Mohammed ibn al-Hasan ibn al-Haitham (Alhazen, c. 965 – c. 1040):

"... not only is every sensation attended by a corresponding change localized in the sense-organ, which demands a certain time, but also, between the stimulation of the organ and consciousness of the perception an interval of time must elapse, corresponding to the transmission of stimulus for some distance along the nerves." (De Boer 2003, pp. 151–152)

Reaction times in general have a long history of study for the insight they provide into the timecourse of cognitive planning (see references in Deary et al. 2011) and *speech* reaction times are widely used in psycholinguistic research. There is, however, a fundamental problem, in that pre-speech articulation is not taken into account in such reaction time measures. Instead, they are all based on acoustics, and while this is recognised as a problem, no solution has been forthcoming to date (Rastle et al. 2005, Kawamoto et al. 2008, Roon 2013).

In this section, we will first review paradigms of speech reaction time experiments. This is followed by a discussion of experiments that are of central interest to this thesis and a summary of the material reviewed here.

2.2.1 Speech initiation experiment paradigms

There are several different reaction time tasks where the reaction time is measured as a vocal or speech response. However, in broad terms these can be divided into classical naming and delayed naming. 'Naming' stands for 'naming out loud', that is, producing a word as prompted by a text, a picture, a recognition task, or by some other means.

A timeline of the speech production stages involved in these two experiment types is shown in Figure 2.6. We see that classical naming includes the lexical processing stage and other neural processes in the latency interval from stimulus onset to response onset. This makes classical naming studies useful in studying the effect of task factors on these processes and these experiments provide data needed in defining psycholinguistic models of speech production such as WEAVER and GODIVA, that were discussed in Section 2.1.7.

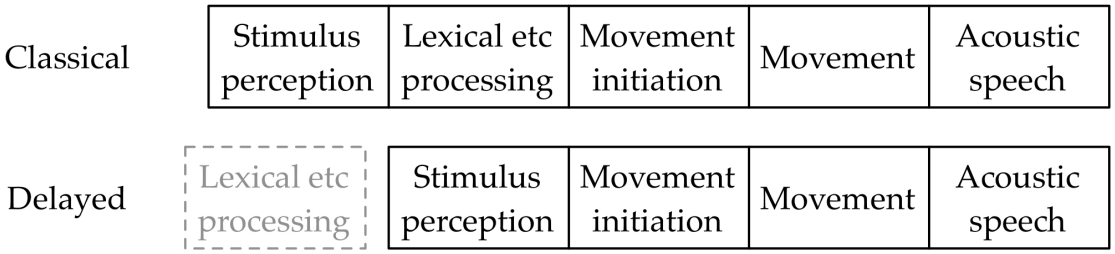

Figure 2.6: Processing stages of classical and delayed naming.

The focus of this thesis makes delayed naming the more useful paradigm, and in the following we will concentrate almost exclusively on variants of delayed naming. Before we do so, let us consider the reaction time data in Table 2.2, which lists reaction time data for mono- and disyllabic words in different naming tasks from studies by Klapp et al. (1973), Klapp and Erwin (1976) and collated by Levelt (1989).

The data shows that structurally more complex words produce longer response latencies when planning is part of the response time (word naming and picture naming), but that this difference becomes very small when either the complexity of the response utterance does not change (categorisation of mono- and disyllabic words with monosyllabic ('yes' and 'no') vocal responses) or the utterance has been planned before the go-signal is given (minimal reaction, that is, delayed naming).

Table 2.2: Mean reaction times for mono- and disyllabic words by Klapp et al. (1973), Klapp and Erwin (1976) and collated by Levelt (1989).

	Word naming	Categorisation	Picture naming	"Minimal reaction"
Monosyllabic (ms)	518.4	695.6	619.3	310.8
Disyllabic (ms)	532.8	697.4	633.3	312.5
Difference (ms)	14.4	1.8	14.0	1.7

Comparing the delayed naming results in Table 2.2 ("Minimal reaction" column) to the articulator specific reaction times in Table 2.1, we see that all of the articulator specific, non-speech reaction times are at least 100 ms shorter than minimal speech reaction times. It remains unclear at this point what causes the

acoustic latency of a delayed naming trial to be so much longer than the simple reaction times of the articulators. The studies reviewed in the next section will shed further light on this and together with the timing results reviewed so far motivate research questions.

There is some evidence for delayed naming reaction times showing effects caused by the content of the response utterance. The naming latency measured as the acoustic speech reaction time increases with the number of syllables in the utterance. While the results from Klapp et al. (1973) show the difference between mono- and disyllabic to be 1.7 ms, Sternberg et al. (1978) report the same difference to be a statistically significant 4.5 ms with a standard error of ±1.3 ms. The results of Sternberg et al. (1978) come from a delayed naming experiment (which is one of several, relevant one is in Section II C in the article) where the participants read wordlists. The lists were composed of 1-4 words. Within condition all words of the list had the same number of syllables (one or two), and the word onsets were phonetically matched across conditions: for example, 'limb' in the monosyllabic condition and 'limit' in the disyllabic condition. The 4.5 ms effect did not change as length of the word list was varied. However, there are methodological concerns, because as mentioned earlier, Sternberg et al. (1978) used a voice key device to determine the speech reaction times. The reliability of these devices will be discussed in the next section.

2.2.2 Acoustic vs. articulatory reaction times

We will start the discussion of acoustic vs. articulatory reaction times by discussing the use of voice key devices in speech research. These devices have received a lot of methodological criticism. The critique has led to a fruitful investigation of how the phonetic identity of the onset of an utterance affects acoustic reaction time. After reviewing the voice key criticism and its results we will discuss studies that fractionise acoustic reaction time into two stages and how that split depends on the methods used.

Voice key

The voice key is a device for automatically measuring vocal or speech reaction times. Voice keys measure vocal reaction time by detecting the acoustic intensity peak associated with the onset of acoustic speech. It was invented by James McKeen Cattell in late 19th century (Cattell 1886). Since its invention, instrumental implementations of a voice key have been used extensively in studying various aspects of the human cognitive systems including, but by no means limited to, speech production and perception (Kapusinski and Rosenquist 1973).

Voice keys are known to have a phonetic bias. Specifically a voice key's results are dependent on the phonetic content of the response utterances in a way that is specific to the model of the voice key used and its settings (Kessler et al. 2002, Rastle and Davis 2002, Yamada and Tamaoka 2003). This problem can be solved to some extent with more careful design of the hardware and/or the relevant algorithms (Tyler et al. 2005, Horbatiuk 2011).

However, it can be argued that the problem is more complex than being more careful about when a voice key should trigger. First, a device-based voice key will always be susceptible to triggering by extraneous sounds (not originating from the speaker) and non-speech sounds (coughs, swallows, other movement sounds). They can be excluded from analysis only by recording the trials and having a human annotator check each trial for correct triggering. Second, if we measure the reaction time based on acoustic output reaching a given threshold level, we are ignoring the fact that participants do not always speak with a constant volume throughout a series of trials. If the voice key is set to trigger at a certain level and the participant for some reason changes the volume of their responses, the voice key will trigger at erroneous times. Third, even if we were to record the speech signal and employ careful manual segmentation to obtain gold standard acoustic speech reaction times, the problem that the speaker moves – that is, initiates speech – usually well before the movements have any acoustic consequences, because articulation (or respiratory movement) always precedes sound production.

James Cattell did not measure only vocal reaction times, but also other

reaction times. His original publication lists three different methods for measuring reaction times: by releasing a telegraph key, by speaking into the voice key device (which he called the sound-key – a less ambiguous term since it is triggered by noise), and by operating a lip key. The last one was a device designed to measure lip articulation, which could be used for measuring speech reaction times as well as simple lip reaction times. It had only one problem in Cattell's opinion: "The only difficulty in the way of using this lip-key is that it is possible for the observer to move his lips before he makes the motion to be registered" (Cattell 1886, page 225). In other words, he could see the lips moving before the lip-key triggered. However, he could not see the tongue and other speech organs moving before any sound was uttered, and thus had no reason not to use the sound-key for measurements. That last point was called into question over a hundred years later as researchers were considering the way speech is produced and specifically trying to separate lexical processing from motor preparation (Rastle et al. 2005, Kawamoto et al. 2008, Mooshammer et al. 2012, Riès et al. 2012; 2014).

Effect of phonetic onset on acoustic latency

Rastle et al. (2005) did an extensive study of the effect of initial phonemes on acoustic reaction time. The study covered all phonotactically legal English *consonantal* onsets in syllables of the three types: /CV/, /CCV/, and /CCCV/. The crucial part of their experiment was to alter the standard delayed naming instruction by asking the participants to remain at an articulatory rest position while they were waiting for the go signal. Before assuming the rest position the participants were allowed to *mentally* prepare and even rehearse speaking the target utterance as much as the participants wanted. After this the participants were asked to produce the by now known utterance in a speeded trial after a randomly delayed go-signal. They "were instructed to produce the target syllable as soon as possible after the tone" (Rastle et al. 2005, page 1087). The results reported by Rastle et al. (2005) show systematic effects of consonant quality on the acoustic onset time. The acoustic onset times are plotted as a

function of the onset consonant duration in Figure 2.7 with colour coding for phonetic identity of the onset consonant.

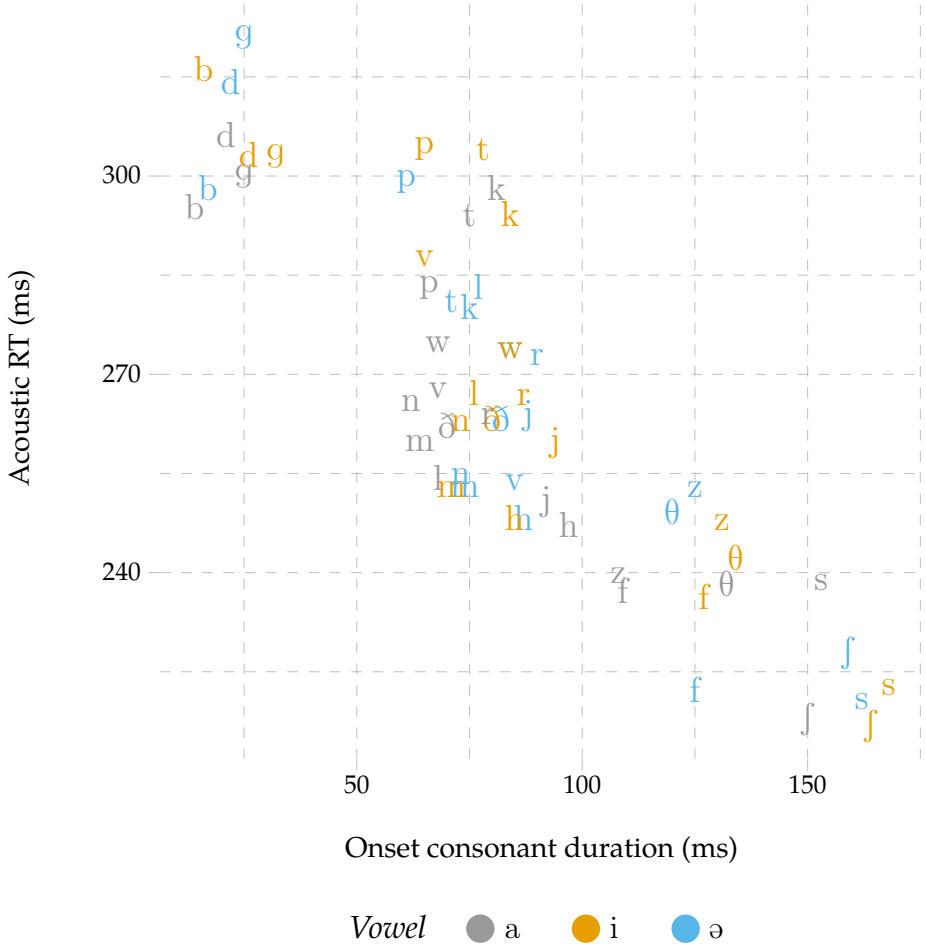

Figure 2.7: Correlation of onset consonant's acoustic duration with acoustic naming latency (reaction time) in delayed naming. This figure shows the data for /CV/ syllables reported by Rastle et al. (2005).

Rastle and colleagues also offer a careful analysis of factors such as voicing, manner and place of articulation, and vowel quality on acoustic reaction time, yet do not discuss in their article the strong inverse correlation of acoustic reaction time and Onset consonant's acoustic Duration (OD) evident in their data (Figure 2.7). They also record and analyse only acoustic data of open syllables, leaving open the role of articulation in the timing of speech initiation and how the timing of initiation events relates to the timing of the whole utterance.

Reaction time fractionation

Kawamoto et al. (2008) analysed a series of reaction time experiments with both delayed naming and classical naming (the latter is not reviewed here). The first experiment compared the effects of different naming delays and consonantal onsets with classical delayed naming instructions. This experiment was recorded in both audio and frontal lip video. The second experiment was recorded only in audio and used a greater range of naming delays with the same instructions. The third experiment was also recorded only in audio, and compared classical delayed naming instructions with the Rastle-type instructions where the participants are asked to remain at rest before they hear the go-signal, and a classical naming task.

Kawamoto et al. (2008) provide systematic measurements of the Articulatory onset to Acoustic onset Interval (AAI) across several task conditions for lip articulations from an experiment using standard delayed naming instructions with audio and frontal lip video. Their other experiments were recorded in audio only. The phonetic materials used were monosyllabic target words (low lexical frequency words in the video taped experiment and non-words in the rest). The words were matched for the number of letters in orthography and number of phonemes in phonological representation across four categories with different onset sounds which were /p/, /t/, /m/, and /n/.

In their first experiment, the only one where lip videos were recorded, Kawamoto et al. (2008) found that if the participants are *not* instructed to remain at rest before they hear the go signal, they prepare the following articulation to a varying degree which is dependent on how long the delay between stimulus presentation and go-signal is. The longer the delay, the greater the articulatory preparation and hence the faster the acoustic reaction time. Their acoustic results (second experiment) also show that the nasal onset words have a shorter acoustic latency than the plosive onset words when the delay is short – the shortest delay they used was 150 ms, but that the difference decreases when the delay is lengthened and is gone when the delay reaches 750 ms. Acoustic results from their third experiment shows that the Rastle instructions preserve

the difference in acoustic latency between plosives and nasals even when long variable delays of 1200-1800 ms were used.

It is unfortunate that Kawamoto et al. (2008) measured articulation only in the standard delayed naming condition and with so few different onset consonants. They also instructed their participants "to respond as quickly and accurately as possible" (Kawamoto et al. 2008, page 353) which may affect the articulation rate of the responses and thus also potentially affect the length of Articulatory to Acoustic onset Interval (AAI).

A different kind of approach to providing a stable reference position was employed by Mooshammer et al. (2012). They did two experiments where they asked the participants produce a prolonged schwa sound ([ə]) while waiting for the go signal. In the first experiment they recorded only audio, and in the second they recorded also articulatory data with Electromagnetic Articulography (EMA). Mooshammer and colleagues used lexical and non lexical words as their stimuli. In addition to the usual singleton and cluster consonant onsets – that is, /C(C(C))V(C)/ words in their case – they included also onsetless tokens – that is, /V(C)/ words.

The acoustic results of consonant onset words recorded by Mooshammer et al. (2012) seem to conform to the inverse correlation pattern of acoustic reaction time and acoustic duration of the onset consonant we have seen in the data of Rastle et al. (2005) (Figure 2.7). Their articulatory results – with their variation of the delayed naming task, namely vocalisation while waiting for the go-signal – show that the locus of the onset dependent delay of consonant onset words is in the AAI. However, their results on vowel onset words are more difficult to fit in the pattern. In acoustics, they pattern most closely with the plosive onset words. If the locus of the inverse correlation is in the AAI, we would expect articulatory onset times not to be affected by the phonetic identity of the utterance onset. However, Mooshammer et al. (2012) report longer articulatory reaction times for vowel onset words than consonant onset words. Unfortunately, besides the difference in the trial task – wait at rest vs. wait while producing [ə] – there are also syllable structure effects in the data that make

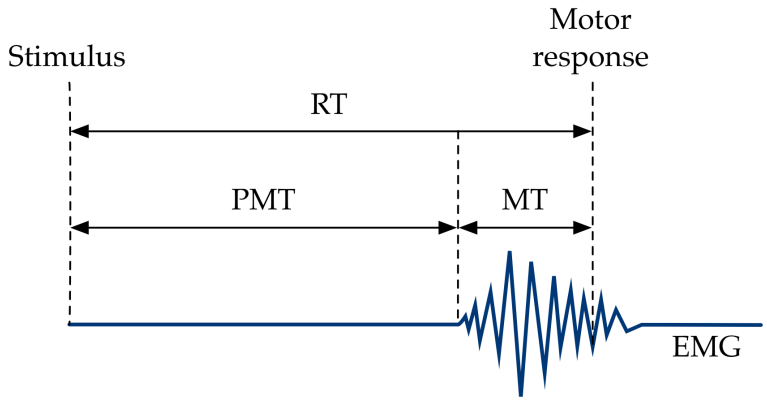

Figure 2.8: Fractionation of finger reaction time (RT) with EMG to pre-motor time (PMT) and motor time (MT). (Image by the author based on an original by van der Linden et al. 2014)

clear interpretation of this question difficult.

A very different approach to separating lexical and other cognitive processing from articulation and acoustic speech production was used by Ries and colleagues (Riès et al. 2012; 2014, van der Linden et al. 2014). They used a measurement paradigm first developed by Possamaï et al. (2002), who used EMG measurements to divide digital reaction times (button presses) to pre-motor time and motor time (Figure 2.8).

The EMG fractionation approach was used by Riès et al. (2012; 2014), van der Linden et al. (2014) to analyse data from a Stroop task (naming colour words that are displayed in either the colour designated by the word or in a different colour). When applied to speech data the motor time is effectively the AAI, with the added difference in this case that as we have seen in Section 2.1.4 neural activation in a muscle precedes muscle activation and movement.

While their approach to separating the articulation phase from the preceding stationary phase looks very effective on the data they have analysed, it has the weakness that to produce a robust division they need to backtrack from the acoustic onset to the preceding EMG pulse – a technique they call the response locked definition of motor time. This is problematic if the phonetic materials being analysed include sounds whose production requires more than one artic-

ulatory gesture before acoustic energy is generated – for example, starting from an open-mouthed position when producing a bilabial voiceless plosive sound.

2.2.3 Summary

Breaking acoustic latency period down to two measurable stages – pre-motor time and motor time or articulatory reaction time and AAI – is the first step to understanding how speech initiation is timed. This is the approach adopted by in the experiments of this thesis.

Drawing together the prediction of Articulatory Phonology (Section 2.1.6) and the inverse correlation pattern found in the data of Rastle et al. (2005) described above, we now see that they seem to be in agreement: the longer the duration of an onset consonant, the earlier it starts. This is despite the fact that the former is based on articulation and the latter on acoustic data. What is still left open is if this pattern is also reflected in the articulatory onset. The studies by Kawamoto et al. (2008) and Mooshammer et al. (2012), lead us to hypothesise that articulatory onset time is not likely to be affected by the phonetic identity of the onset consonant, but without new data we can not know this with certainty.

There are certain problems even with tasks as simple as these as pointed out by Kawamoto et al. (2008). Let us consider, for example, the production of the syllable [pa] as part of a delayed naming experiment. If we wanted to know the motor time for producing the syllable we should get our participant to stay at some sort of rest position before giving them the signal to produce the syllable. If, on the other hand, we were interested in knowing the minimal reaction time of lip opening we might give the opposite instruction and tell the participant to be ready to produce the syllable with the shortest possible reaction time. This should cause them to go through all the preparatory articulation leading up to the release of the plosive [p] and then hold that position until a go-signal is given. So, care needs to be taken in considering the instructions and tasks that are given to the participants.

The experiments in this thesis focus on tongue movement. This decision is based on the tongue being the fastest articulator to respond in simple reaction

time tasks (Section 2.1.4), and the fact that it is involved in the production of most speech sounds. While lips also show fast minimal response times, they were ruled out, because it is possible to initiate speech articulation inside the mouth without opening the lips. Furthermore, a number of ways of measuring tongue articulation are available. The next section reviews their properties and justifies the use of Ultrasound Tongue Imaging (UTI) as the main recording method in this thesis.

2.3 Articulatory measurement methods

Most speech production research relies on acoustic data, but acoustic data provides only an indirect view of the process itself. Since we are trying to understand speech initiation and especially the silent articulation that takes place before the acoustic onset, we will need to record and analyse articulatory data. This section lays out the rationale for choosing ultrasound as the main articulatory recording method for this thesis.

Before we review potential recording methods, it should be noted that all of the methods reviewed here produce time dependent – that is, dynamic – data from the tongue. Static methods – for example, plaster casts (Chiba and Kajiyama 1941, Ladefoged et al. 1971) and 3D Magnetic Resonance Imaging (MRI) (Ericsdotter 2005, Palo 2011) – while useful in other studies of speech production, can not be used in studying speech initiation, because speech initiation is not a static phenomenon. The following sections review the most relevant articulatory recording methods. UTI is reviewed last, because it is the main articulatory recording method in this thesis, and thus, receives the most careful review.

2.3.1 Electropalatography (EPG)

Electropalatography (EPG) is a method for dynamically measuring tongue-palate contact (Hardcastle 1972). The measurement uses a thin artificial palate, which contains the pressure sensors, that detect tongue contact. The contact

pattern can be sampled with a high frequency by attached analysis equipment. An example of an artificial palate is shown in Figure 2.9.

An EPG palate can affect speech patterns and it is recommended, that participants get used to wearing the artificial palate before measurements are made. In addition, the palates need to be custom fitted and the manufacturing process is not instantaneous and is not cheap. This means that preferably the same participants should be re-recruited over and over again to get maximal benefit from invested time and resources.

In terms of acquired data, EPG is limited to recording only tongue-palate contact pattern. No information is gained on the tongue-palate distance or movements of other articulators. On the other hand, EPG is safe and has a fast sampling rate – usually either 100 Hz or 200 Hz. The spatial sampling is also quite dense with electrode numbers ranging from 60 to almost 100.

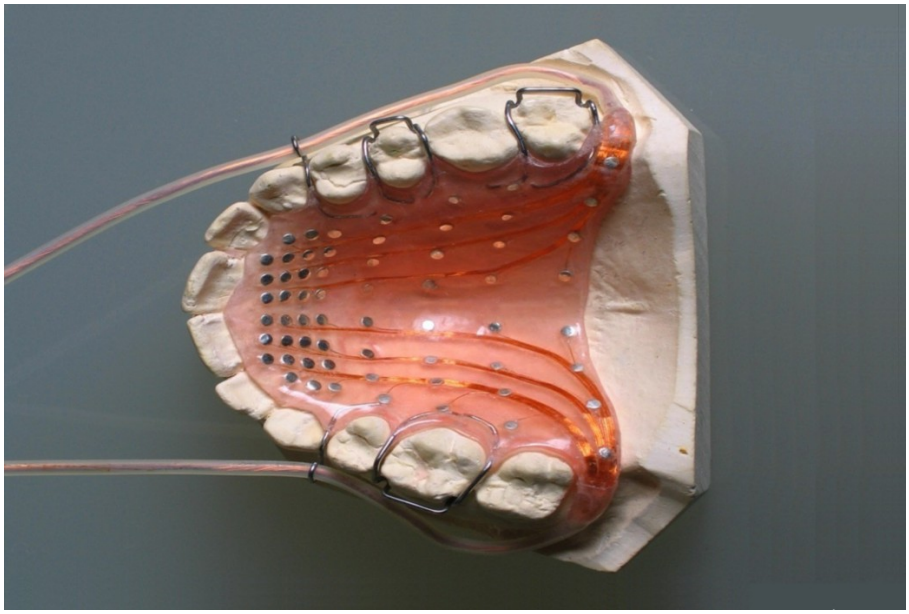

Figure 2.9: An artificial EPG palate mounted on a dental cast. The round dots on the artificial palate are the contact sensors and their wires can be seen exiting the palate behind the molars. When the EPG palate is mounted in place on a participant's palate, the wires are routed out of the corners of the mouth. Please note, that different manufacturers of EPG equipment use different sensor configurations. (Image courtesy of Professor Alan Wrench.)

EPG is a well-used and well understood method, that produces results, which are readily comparable with other studies. EPG data has established analysis methods (Hardcastle et al. 1991) that are potentially useful also in analysing pre-speech articulation. It has, for example, been used to analyse tongue palatal contact patterns in speech preparation by Fuchs and Ünal-Logacev (2017). On the downside, EPG has a limited view of the tongue movements as it only registers palate contact and therefore has no direct information on tongue movement that does not reach the palate.

2.3.2 Optopalatography (OPG)

The use of optical distance sensing equipment to study intra-oral articulation was first suggested by C-K. and Wang (1978) and later developed under the name Glossometer – a device that used LED/photodiode pairs on the midsagittal line of a false palate – by Fletcher et al. (1991). Optopalatography combines the idea of the Glossometer with that of EPG and aims to broaden data gained from palatographic measurements by providing tongue-palate distance data as well as readings on the firmness of tongue-palate contact. A working prototype Optopalatography (OPG) device was developed by Wrench et al. (1996; 1997; 1998).

Optopalatography (OPG) uses laser distance measurements to provide a 3D image of the tongue surface within the oral cavity. Instead of mounting electric contact sensors on the artificial palate to provide tongue contact information, OPG uses optical fibres mounted on the artificial palate to provide tongue distance information. The fibres work in pairs: One fibre carries the light produced by the OPG unit to the palate and sends it towards the tongue and the other catches returning light and transfers it back to the OPG unit for analysis.

The most important limitations for OPG system design are caused by the qualities of the optical fibres. Fibre diameter limits the number of possible measurement points as the size of fibre bunch, that can be comfortably relayed out from the corners of a participant's mouth is quite limited. On the other hand, fibres with a very small diameter can not be used as they have to be bent

90° within the thickness of the artificial palate.

Another important limitation is the signal-to-noise ratio, which is affected by light source strength, fibre qualities and possibly by the proximity of the sensor/source pairs in the artificial palate. The proximity problem can be avoided by operating the points in sequence with only one sensor/source pair active at any given time.

When compared with EPG, OPG has two additional good qualities: It produces tongue-palate distance measurements and can be used for force measurements, since all of the light is not actually reflected by the surface of the tongue, but also from within it. Thus, the firmer the contact – and the greater the force – the stronger the reflection. In addition, when compared with the Glossometer, OPG clearly facilitates more measurement points.

The original prototype Wrench et al. (1998) had 16 sensors built from 0.5 mm plastic optical fibres. The system had a measurement range of 20 mm and sample rate of 100 Hz. The light sources were infra-red LED sources, and they were used in sequence.

In the current context OPG has three short comings. First, and most importantly, it is not yet a mature method. While the original OPG development project never got further than the prototyping stage, there has been recent work on an improved system (Stone and Birkholz 2017). The new project, however, is still at the prototype stage. Second, like EPG, OPG requires a palate fitted to each individual speaker, making it a somewhat expensive method to use. Third, even though OPG would have good time resolution, the imaging area is limited to the front part of the tongue.

2.3.3 Electromagnetic Articulography (EMA)

Electromagnetic Articulography (EMA) is a point tracking method capable of tracking several tagged points on the face, teeth and tongue of a participant. The tracking is based on measuring changing magnetic fields at the tracked points with small receiver sensor coils which are connected by wires to the measurement unit. In older systems, the measurements are limited to the mid-

sagittal plane, while more recent systems are capable of observing movement in all three dimensions (Perkell and Oka 1980, Schönle et al. 1987, Hoole and Zierdt 2010). The fields are produced by transmitter coils, each of which has its own field frequency. The fields of the transmitter coils induce small currents in the receiver coils and the measurement apparatus records these currents for each receiver coil.

The spatial tracking in EMA works by estimating from the measured current strength the distance from the receiver coil to the transmitter coil. The process involves estimating the local field strength at the receiver from the current strength and computing the distance thereafter from the field strength. Problems arise if the estimated field strength is incorrect, which, in turn, is caused by misalignment of the receiver. To work well, old mid-sagittal or Two Dimensional (2D) EMA devices required, that the receiver and transmitter coil's main axis be parallel. If this was not the case, the distance would be overestimated.

After more than a decade of development from early 1970's (Hixon 1971, Lance and van der Giet 1974, Sonoda 1974), EMA systems have been commercially available from mid-1980's starting with 2D systems such as the Movetrack by Branderud (1985). More recently, EMA devices which are not restricted to the mid-sagittal plane have matured and become the standard (Zierdt et al. 2000, Hoole et al. 2003). Figure 6.2 shows the AG500 system in use. It is a system that uses six transmitting coils attached to the plastic frame surrounding the participant. With each transmitter transmitting a different frequency the system is able to track each receiver in 3D and provide two rotational coordinates for them as well.

Nowadays, there are commercially available EMA systems which can record between 8 and 24 individual tracked points (Savariaux et al. 2017). The sample rates of modern EMA systems go up to 1250 Hz, but are usually used at lower effective sampling frequencies to reduce noise in the data. A maximum spatial resolution of 0.5 mm and reported maximum median error of 2 mm (Yunusova et al. 2009, Kroos 2012, Savariaux et al. 2017). In current, systems

there is a trade off between portability and precision: the Northern Digital systems are better for portability, while the Carstens AG501 is the best choice in terms of precision according to an independent study by Savariaux et al. (2017).

The greatest shortcomings of EMA lie in what can be measured: EMA tracks only points and the method cannot be used very deep inside the vocal tract as the coils and wires involved would trigger the participant's gag reflex. Furthermore, the receiver coils have to be connected with wires to an analysis unit. This means, that when measuring intra-oral articulation, there will be wires passing into the participant's mouth. In addition, the articulation's naturalness can be affected by the receiver coils themselves – especially if the coils are placed too close to the tongue tip, or there is additional equipment placed in the speaker's mouth (Hoole and Nguyen 1999).

Nevertheless, EMA is a popular method for measuring articulation. The main causes for its popularity are its safety, its good time resolution and the fact that after the initial cost of acquiring the equipment it has a fairly low operating cost. In addition, it can be used simultaneously with other methods like EPG (Engwall 2000b) or ultrasound and electroglottography (Grimaldi et al. 2008). It has also been used for studying speech initiation by Mooshammer et al. (2012).

It is also used in this thesis, but only as the secondary measurement method. As we have seen in Section 2.1.4, there is a lag between neural activation signals arriving at the muscles of an articulator and that articulator actually moving. EMA is a fleshpoint tracking system and can only record the movement of the articulator, but will have no access to the internal processes of a muscle that lead to overt movement. There is also a potential problem in the use of EMA being limited to the externally visible articulators and the upper parts of the vocal tract.

2.3.4 Magnetic Resonance Imaging (MRI)

Magnetic Resonance Imaging (MRI) utilises magnetic resonance of hydrogen nuclei to produce a tomographic image of the object or tissues being studied. MRI is more properly called nuclear magnetic resonance imaging – NMR imag-

ing for short. However, "nuclear" has been dropped out of use in medical contexts and hence articulation study contexts. This was done since the use of "nuclear" falsely suggested to patients and participants, that the method would use ionizing radiation.

Instead, the imaging procedure involves radio frequency oscillations of the magnetic field to produce a measurable echo field from the hydrogen nuclei. To produce meaningful data the spins of these nuclei have to be directionally aligned. This is accomplished by the application of a very strong static magnetic field. The gathered data is then processed with mathematical inversion methods to produce the tomogram.

MRI does have some drawbacks, most of which can be overcome with careful planning or by the use of state-of-the-art equipment. First and foremost it should be mentioned, that the participant's safety has to be considered carefully. As the method involves very strong alternating magnetic fields the patient should not have any metallic implants or extraneous metallic material within his or her body. The metallic material may start to heat up and cause injuries or, in the case of small particles, it may even move back and forth with the changes in the magnetic field, and thus damage the tissue it moves through. These same restrictions apply to any extra equipment used within the magnetically shielded scanning room. While the no-metal restriction is not absolute it has been found that, for example, the wires for a normal microphone tend to pick up noise from the oscillating magnetic field, which often renders the recordings useless.

Another problem is the acoustic noise level within the scanning room. Modern MRI scanners produce noise through the magnetostrictive effect during rapid field changes which occur during the scanning procedure. This poses another problem for simultaneous sound recording. (See below in this section 2.3.4 for possible solutions to this problem.)

Finally, there are three problems with the imaging itself. These are the typically long acquisition times needed for full 3D scans of the vocal tract, the occasionally poor air-tissue contrast and the fact that MRI does not produce practically any signal from calcified structures such as bones or teeth. The

acquisition time and the air-tissue contrast problems are disappearing with advances in equipment technology and imaging protocols. In contrast, the problem with bones and teeth is inherent to the way MRI works. There is very little hydrogen in the calcified tissues, and since the method images tissues by detecting the electro-magnetic echoes from hydrogen nuclei, calcified tissues are rendered practically invisible in MRI.

However, when used correctly MRI does not cause any known health risks. Indeed, when compared with the cineradiographic techniques which have been very popular until late 1980's, MRI has two very important advantages. Despite its original name, it does not use ionizing radiation. It can, therefore, be used for large corpus studies and for repeated measurements with the same participant. In addition, MRI is a volumetric imaging technique, which is able to produce images with good spatial resolution at least with modern equipment. A further bonus is the possibility of choosing the angle and plane of the tomograms freely when capturing 2D images.

Real-time or fast MRI differs from regular or static MRI in terms of imaged area and temporal resolution. The imaged area is always smaller in fast MRI than in static MRI - usually it is restricted to a single mid-sagittal slice of the vocal tract. With this reduction comes the advantage of speed: Instead of using several seconds or even tens of seconds to image the whole vocal tract, tomogram frames of the vocal tract can be captured at rates of up to 83frames per second (fps) (Lingala et al. 2017).

Sound recording during MRI

Sound recording is an essential part of data gathering for analysing the timing of articulation in relation to acoustic speech. Sound recording during MRI poses some challenges. In particular, if speech should be recorded during an MRI sequence, the over all conditions are quite difficult (Demolin et al. 2000, Palo 2011). First, there is a high level of noise from the MRI device itself during a scan. Second, only very small amounts of ferromagnetic material maybe brought into the scanner room, because of the very strong magnetic field generated by the

scanner. Even the small amounts of ferromagnetic material need to be kept away from the scanner itself lest they be caught in the static magnetic field of the scanner. Third, any electronics used in the scanner room should be shielded from the powerful alternating electromagnetic field that the scanner produces during a scan.

The high level of noise means, that the sound recording system has to either be able to separate the background noise from the speech and be directional enough not to record too much of the noise in the first place. The static magnetic field makes the use of regular microphones difficult and the alternating field means that unshielded cables are likely to pick up a lot of interference during a scan.

As demonstrated, by Engwall and Badin (1999), sound can be recorded before and after each scan with a regular microphone, if a screened cable is used and the microphone is situated far enough from the scanner. This, however, means that the microphone will be far from the participant. Since no useful sound can be recorded during the scan, the speech productions need to be static. This means that such an arrangement cannot be used to study speech initiation.

Two possible solutions for the sound recording problem are optical microphones or a regular microphone used outside of the scanner's magnetic field, in combination with sound carrier tubes that sample the speech of the participant close to their mouth. To work well the latter solution needs to include some form of passive sound capturing equipment to capture the speech of the participant acoustically.

An optical differential microphone system is available commercially and has been used by for example Ericsdotter (2005). The system consists of a light reflecting membrane, whose vibration was measured optically. The optical signal was then transduced to an electrical one at a safe distance from the MRI scanner and after amplification it was recorded with a DAT recorder. After a post-recording noise filtering stage a vowel's fundamental frequency and strongest formants could be measured.

As for capturing the participant's speech with passive acoustic components, the use of a pneumatic mask for this purpose has been proposed by Demolin et al. (1997). An actually working system using a sound collector that rests in front of the supine participant's face suspended by the registration coil has been constructed and used by Lukkari et al. (2007), Malinen and Palo (2009), Palo (2011). The system uses wooden and plastic parts within the scanners magnetic field to capture two directional sound signals into tubes that transport the sound to a shielded microphone array outside of the MRI scanner's field.

MRI in speech initiation studies

MRI is a good candidate for an articulatory recording method in studying pre-speech. There are two problems with its use though. First, while the sound recording problem does have available solutions, the methods are not yet capable of reliably removing all of the scanning noise without interfering with the speech sounds recorded. This would make reliable identification of acoustic onsets impossible. Second, there are restrictions placed on access to the facilities by the medical institutions that own most of the scanners, and scanning time is usually expensive. This means that scanning time is not easily available for purposes such as piloting and additional funding would be needed for recording substantial amounts of data.

2.3.5 X-rays and related methods

X-ray based methods such as cineradiography, X-ray microbeam, and Computed tomography have traditionally been popular, but are excluded from use in this thesis for ethical reasons. In recent, years their use has been limited to case studies, such as the one reported by Vampola et al. (2011), where the sole participant was a willing volunteer and as one of the authors fully aware of the risks of the method used. But since these are fast imaging methods, a review of their properties is included here for completeness.

Cineradiography

Until late 1980's, cineradiography was the method of choice for studying speech movements (Dart 1987). Cineradiography uses ordinary X-ray apparatus. Instead of taking still pictures the apparatus is used to record a cinematic sequence of X-ray pictures with a specialist camera. While the method can produce interesting data, it has some serious drawbacks.

The most important drawback and the reason for cineradiography's decrease in popularity is radiation. As awareness of the long term effects of ionising radiation increased, the regulations on the maximum allowed dose became stricter. This can be considered only a good thing, since the safety of experimental participants should always be a prime consideration. It does, however, mean that gathering a large corpus of data with cineradiography is out of the question.

Another significant drawback is the fact that cineradiography is a transillumination method. It flattens the imaged object in one dimension. Since, the data is customarily recorded only in a sagittal orientation, lateral structures are superimposed on each other. This leaves any asymmetries out of the data as well as making it hard to judge changes in the coronal direction such as grooving. This adds to the difficulty of capturing the articulatory movements of certain sounds such as [l]. Moreover, as the teeth and certain parts of the tongue are quite often projected on top of each other in a sagittal projection, it is sometimes hard to make out the correct contour of the tongue from the images.

On the positive side, the recording position in cineradiography is usually very natural and obviously the method is dynamic in nature. It should be noted, that while new data cannot be recorded, data from old studies is often available in a useful form (Munhall et al. (1995)). From the point of view of this thesis, the fact that new data may not be recorded, renders this method uninteresting.

X-ray microbeam

X-ray microbeam is a point tracking method developed originally by Kiritani et al. (1975). It is based on ordinary X-ray technology. However, instead of imaging the whole vocal tract only small lead or gold pellets are tracked. The tracking is performed with a very small X-ray beam. The beam's penetration of the tissues and the pellets is registered with a digital detector array on the other side of the object.

Before tracking begins, the pellets are first located by scanning the whole image area. In contrast, during tracking only a neighbourhood around the pellets' last locations is sampled. This reduces the amount of radiation exposure greatly. The only problem with this is that specifically tissues around the pellet and its projection absorb most of the radiation. Even so, the amount stays quite small even during a recording of a comparatively large corpus of data.

The early system had a spatial resolution of 1mm and an effective sample rate of 100 Hz Kiritani et al. (1975). The pellet diameter was originally 3 mm, but dropped to 2.5 mm as the material changed from lead to gold Fujimura (1991).

While originally X-ray microbeam was considered a good way of lowering the radiation dose absorbed by a participant, it went out of use when it was realised that the local dose around the target pellets was still considerable, and the rapid development of EMA methods replaced it as a flesh point tracking method in speech studies.

Computed Tomography

Computed tomography is a 3D imaging method which is based on X-ray imaging. An intermediate step in the historical development of computed tomography was ordinary tomography. It was a method of producing tomograms - pictures of a slice through the imaged object - on regular X-ray film. X-ray imaging and ordinary tomography have been used in articulatory studies and are discussed in Section 2.3.5.

Like ordinary tomography, computed tomography produces 2D slices of

the participant. The slices are the result of computing the tissue density by solving the inversion problem for X-ray absorption data. The absorption data is gathered by rotating an X-ray source around the participant and registering the amount of radiation passing through the participant on sensors positioned opposite to the X-ray source. 3D images are produced by obtaining several parallel slices of the same participant.

While early instruments were naturally slower, more modern ones have a fairly fast slice acquisition time. Even so, they have been far from being instantaneous in acquiring a 3D image. This is changing with the introduction of multislice-computed tomography, which can at present acquire up to 16 slices in one pass.

Computed tomography produces images with a good contrast between air and bodily tissues in general. For the purposes of imaging the vocal tract this is especially good as the walls of the vocal tract will be clearly defined and easy to extract.

Moreover, computed tomography has good resolution. Modern devices are capable of producing a slice thickness of 1mm and plane pixel size of 0.2mm. These are, however, parameters for clinical use of computed tomography. If used for non-medical purposes - such as speech production studies, it is necessary to be stricter about the radiation doses, and thus accept lower resolutions.

The main flaw of computed tomography is its use of X-rays. There will inevitably be an amount of ionising radiation absorbed by the participant. The obvious need to keep the radiation dose for the participant in acceptable limits restricts the number of images that can be acquired from one participant. Out of all of the methods that use X-rays, computed tomography is still sometimes used in a limited capacity in speech studies (Vampola et al. 2011).

2.3.6 Ultrasound tongue imaging (UTI)

The ultrasound images we are used to seeing of infants in the womb usually come in the form of a fan shaped greyscale image. They are in fact formed by an application of sonar. The same imaging systems can be used to image tissue-

air boundaries and tissue internal structures in the tongue and surrounding anatomical structures. Ultrasound imaging works by sending ultrasound pulses into the imaged tissues and listening for echoes coming back. The echoes are then used to construct an image of the structures that reflected the ultrasound pulses by interpreting the echo delay as distance from the probe and equating echo strength as physical distinctness of the reflecting structure.

Ultrasound is usually used in B-mode (Brightness mode) for speech research (Stone 2005). B-mode generates a sequence of images of relative brightness information based on the echoes of radio frequency ultrasound pulses. The pulses are in the range of 3-9 MHz, and they are generated and received by the ultrasound transducer, which is situated at the head of the ultrasound probe (Figure 2.10). Medical ultrasound devices are generally designed for imaging static structures rather than recording movement.

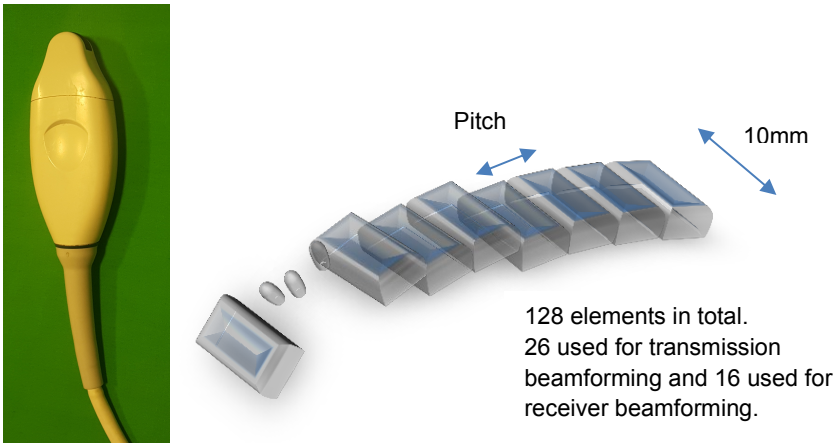

Figure 2.10: Left: the microconvex ultrasound probe used at Queen Margaret University (picture courtesy of Dr Sonja Schaeffler). The transducer is housed in the top of the probe behind the strip of darker material. Right: a schematic of the composition of the transducer array at the head of the probe (Image by Wrench and Scobbie 2016, , used with permission.).

In contrast, speech researchers are mainly interested in imaging the tongue or the larynx in motion. This means, that there are a number of issues in selecting the hardware and choosing scan settings that need to be taken into account to avoid artefacts and errors in the data (Stone 2005, Wrench and Scobbie 2006).

The settings need to also be optimised to ensure that as many as possible of the relevant structures are imaged with a high enough frame rate. For this reason the probe in Figure 2.10 is rarely used with its full field of view – with all transducer elements in use the probe has a field of view of 150 degrees – rather it is usually used with a smaller field of view – imaging a sector of less than 150 degrees – to increase the frame rate.

Ultrasound imaging – or sonar – is a form of echolocation. Imaged structures are detected by bouncing an ultra-high frequency (radio frequency to be exact) sound beam off them. While somewhat similar in principle, ultrasound imaging is not laser distance measurement. Instead of the very coherent and tight beam that can be achieved with laser physics, ultrasound has to use a less ideal beam formed by firing several elements of the piezo-electric transducer array in a probe (Figures 2.10 and 2.11).

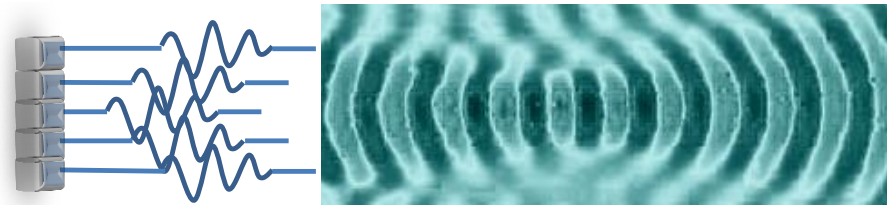

Figure 2.11: A single ultrasonic beam or scanline is generated by firing a number of probe array elements in a sequence that produces a constructive interference pattern along the desired direction (Image by Wrench and Scobbie 2016, , used with permission.).

UTI commonly uses ultrasound probes which produce a fan shaped image of the tongue. Ordinary or interpolated ultrasound data refers to the form usually displayed by ultrasound imaging systems as seen in Figure 2.12 c). The fan image of the ordinary ultrasound data is produced by interpolation between the actual raw data points returned by the probe as it images the tissues (Figure 2.12 a). The precise method of interpolation differs from one system to the next and can potentially depend on the version of software used. The raw data points are distributed along radial scanlines with the number of scanlines and the number of data points imaged along each scanline depending

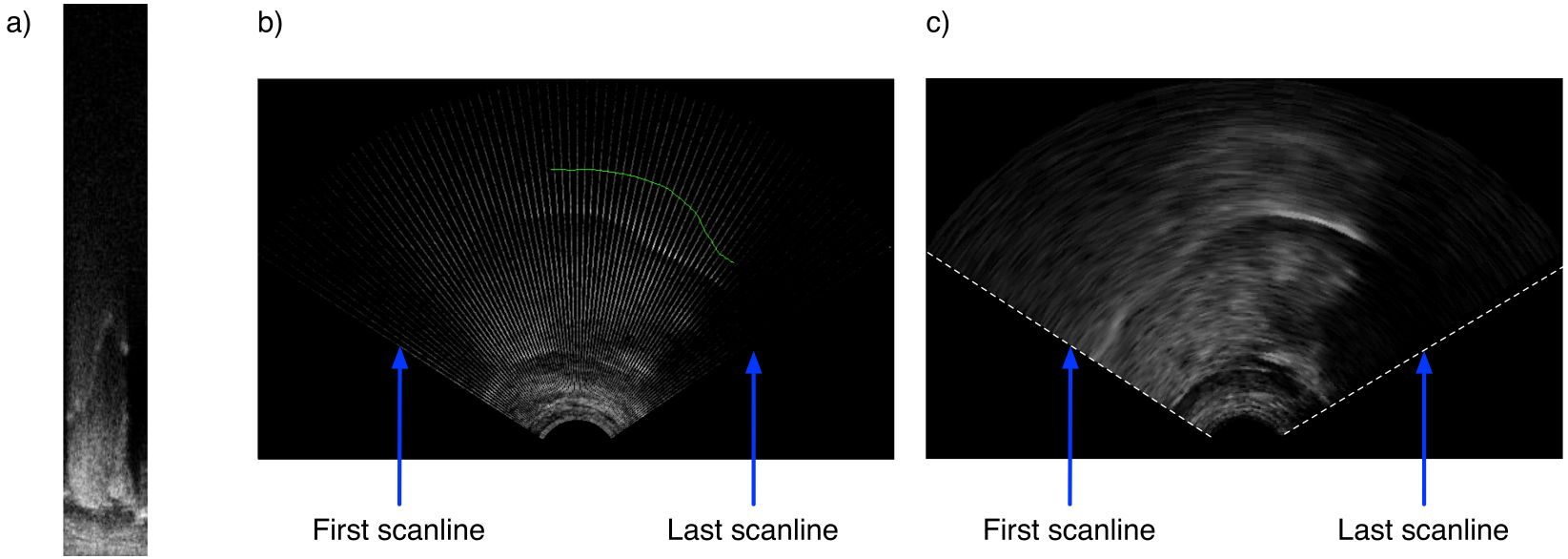

Figure 2.12: a) Raw (uninterpolated) probe return data. Each beam or scanline is represented by one column of pixels. b) Scanline data spread on a fan corresponding to the acquisition angle of each scanline. (Green line is a palate trace) (The ultrasound frame and palate trace by Wrench and Scobbie 2016, , used with permission, arrows and scanline annotations by the author) c) Fully interpolated ultrasound image.

on the setup of the ultrasound system (Figure 2.12 b).

An ultrasound image is produced on the basis of several ultrasound pings (single flashes) sent into the tissues being imaged. Each ping produces echoes as the sound pulse travels across boundaries between and within tissues. These echoes are used in reconstructing the tissue structure that the ping travelled through – to the extent that the echoes actually reflected back to the transducer. There are three main reasons why an echo might fail to return:

1. There is no echo. If the pulse encounters a medium with a significantly higher speed of sound – such as bone – this medium sucks in almost all of the pulse's energy with minimal or no reflections/echoes bouncing back towards the transducer.
2. The echo signal was weakened too much by travelling too far in the tissue medium and has become masked by background noise. The soft tissues of the body (and therefore of the tongue) attenuate ultrasound pulses a lot as they pass through these tissues causing more and more severe signal decay as the pulse travels further into the tissue. This causes a practical limit on the depth of an ultrasound scan and limits the visibility of the tongue contour in people with large tongues and in articulations such as [k] and [g] where the tongue moves towards the velar region and consequently further away from the probe.
3. The echo went somewhere else. If a surface within the tissues is not perpendicular to the direction of the pulse's travel the echoes will not be reflected directly back toward the transducer, but will instead scatter into the surrounding tissues. This happens frequently – especially with inner structures of the tongue (mainly muscle fibres).

In Figure 2.12 c) we see that there is a lot of image noise or fuzziness in a single UTI frame. This noise – often called speckle noise – moves from one frame to the next confounding especially untrained observer's' perception of where the tissue or air-tissue boundaries are in the image. Depending on the system, this speckle maybe mainly actual noise resulting from unideal qualities of the imaging equipment. However, in a high-end system, such as the facility

at Queen Margaret University, it will be mainly the result of individual muscle fibres changing state (muscle fibre activation is discussed in Section 2.1.4 and its relation to ultrasound data later in this section). This is true for both pixels below the tongue surface (that is, where the muscles actually are) and *above* it. The pixels above the tongue contour are a result of an ultrasound beam getting reflected around inside the tongue so that it takes a longer time to travel back to the transducer then a direct echo returning from the tongue surface. Since the ultrasound system interprets the echo return delay interval directly as distance from the probe, these echoes are interpreted to have originated from above the tongue.

As can be seen in Figure 2.13 the depth resolution of ultrasound is good, but small structures tend to smear across the imaging fan. The properties of the beam can be adjusted to give a better focus at a desired imaging depth in the reconstructed image, but due to the way the beam is formed, this is always a trade off against worse resolution at other imaging depths.

There are a number of procedures that can be carried out as part of or during the recording of UTI data to make the data more readily interpretable. First, we need to know where the probe is in relation to the participant's anatomy. This can be achieved by holding the probe in a stable manner in relation to the head of the participant. In clinical use and when working with small children as participants, this is often achieved by either the participant or the therapist holding the probe in their hand. However, in laboratory experiments involving adult speakers the probe is usually stabilised mechanically in some way (Stone 2005) or partially stabilised with head-probe-position correction provided for example by optical tracking (Whalen et al. 2005, Wilson 2006). In Queen Margaret University's ultrasound laboratory the method of choice is the headset shown in Figure 2.14 and provided by Articulate Instruments Ltd (Articulate Instruments Ltd 2008).

When using a mechanically stabilised probe to guarantee a reasonably constant probe-to-head position, two procedures can be employed to give information on the relationship of the images to the general head anatomy of

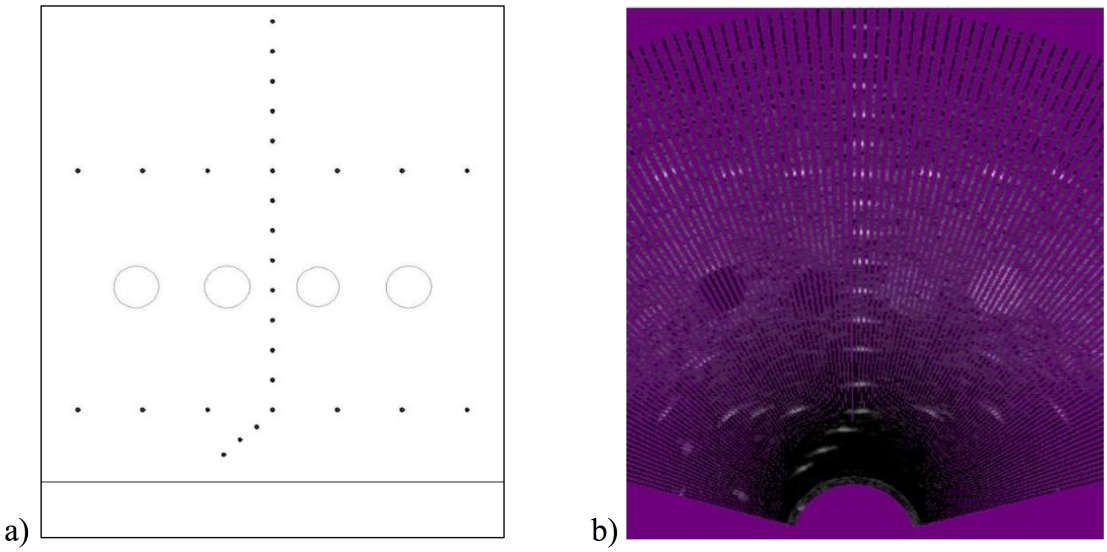

Figure 2.13: Evaluation of the imaging properties of the ultrasound system used in this thesis: a) Phantom target structures b) the echo returns from multiple beams with an angular separation of 1.2° (124 beams over 150◦ field of view). The point objects at 6cm depth appear on approximately 5 adjacent scanlines. This means a point target is smeared by ± 2.4° (Images by Wrench and Scobbie 2016, , used with permission.).

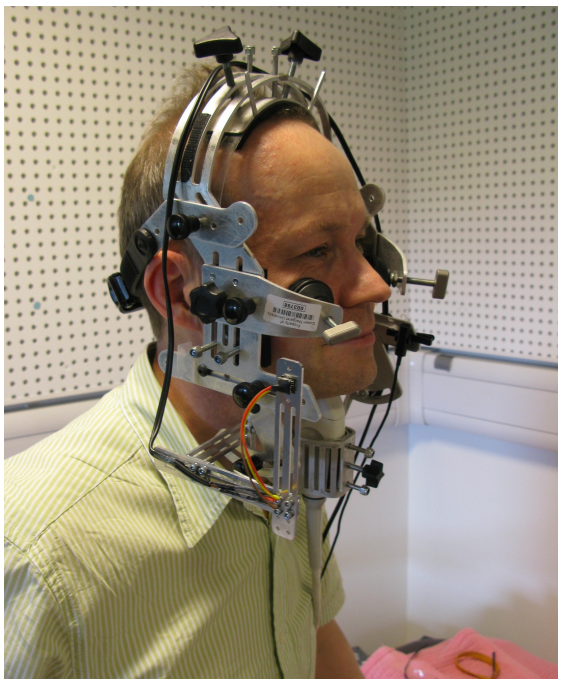

Figure 2.14: Ultrasound headset. (Picture courtesy of Professor Alan Wrench.)

the participant. First, a bite plate can be used to find out the orientation of the UTI images in relation to the participant's occlusal plane (Scobbie et al. 2011). Second, the participant's palate position in the images can be acquired by recording a water swallow and tracing the hard palate as the upper boundary of the moving water bolus (Stone 2005).

A completely upright (axial in relation to the participant's anatomy) probe position is very rarely optimal in UTI. Figure 2.15 shows a typical ultrasound frame from the data corpus of this thesis matched to an anatomical trace obtained from MRI. Regardless of the fact that 'up' in an ultrasound frame is rarely 'up' in relation to the participant's anatomy the images aren't usually rotated for display purposes.

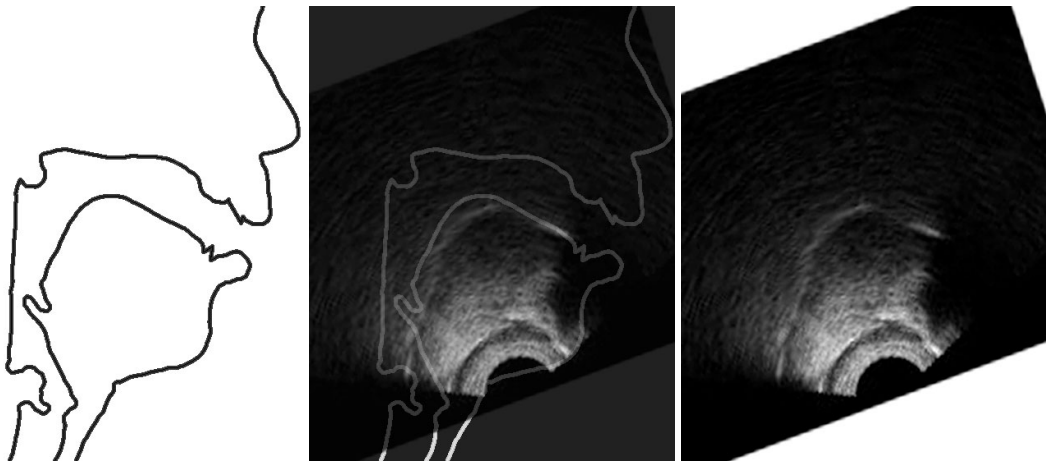

Figure 2.15: General relationship of ultrasound data to head anatomy and orientation shown by overlaying (middle pane) an ultrasound frame (right panel) on anatomical contours extracted from a magnetic resonance image (left panel). The image is based on two different participants and is only provided to demonstrate a typical tongue ultrasound image's general orientation. The specific orientation changes between individuals and recording sessions.

UTI in speech initiation studies

Ultrasound has many desirable qualities when studying speech production. Modern devices have good time resolution – devices used in speech studies regularly have frame rates of 60-120 fps. This is combined with good spatial

resolution (See Figure 2.13 above). High end devices, such as the one at Queen Margaret University, even offer flexibility by letting the user adjust the spatial resolution, time resolution, and imaged area (improving one comes with the cost of degrading one or both of the others) to find the best balance of the three.

The main flaw of ultrasound is that it can not image structures through an air gap nor through bone. So while most of the tongue can be imaged (and indeed more of the tongue then with, for example EMA or EPG), the tongue tip is often behind an air gap or behind the mandible shadow. Furthermore, the opposite wall of the vocal tract will be visible in ultrasound images only when the tongue is in direct contact with it. Nevertheless, having good quality data from most of the length of the tongue makes ultrasound a good choice for speech initiation studies.

What might be considered another undesirable quality is the presence of the speckle noise. The speckle noise, however, turns out to be an asset that we can exploit. As mentioned above, it is the result of changes in the state of muscle fibres. In other words it contains information about muscle activation. This has also been verified in a number of studies. Koppenhaver et al. (2009) provide a general review of studies on how muscle activation and movement registers in ultrasound with comparisons to EMG data. To give an example, Vasseljen et al. (2009) studied activation onset in abdominal muscles with ultrasound imaging and fine-wire intramuscular EMG. They extracted activation onset from M-mode ultrasound and related to it to neural activations derived from EMG data that was recorded from the same muscle. Their results show a strong correlation between ultrasound onsets and EMG onsets.

2.3.7 Summary

There are many advanced methods for collecting tongue data. Most of them have qualities that either preclude their use or make them less desirable than others. EPG offers only tongue-palate contact information which is not enough to reliably detect movement onset. OPG would remedy this, but the method is not available yet, and in any case, it will be only able to image the front part

of the tongue. MRI would offer good coverage of the articulatory organs with good time resolution (when cutting edge hardware is available), but suffers from problems in sound recording. X-ray based methods can not be used in extensive data collection because of participant safety concerns.

EMA is a fleshpoint tracking method that can be used on multiple articulators simultaneously. Its main flaw is that it uses sensors that can only be attached to fairly frontal portions of the tongue. It is used as the second articulatory data collection method in this thesis.

In the current context UTI is a good choice. Its main drawback is that using a headset for holding the ultrasound probe affects the speaker's articulation. However, UTI does offer good time and space resolution, and additionally, it offers the possibility of detecting internal muscle activation in the tongue with suitably developed methods. The next section will justify the choice of data analysis methods used in this thesis.

2.4 Data interpretation and analysis methods

Analysing articulatory data often requires a lot of experience and time. This poses an efficiency problem in many projects and limits the amount of data that can be analysed. There is also an issue in that, while we have established, standard ways of visualising audio data in the time domain (waveforms and spectrograms), no comparable standardised methods exist for video data such as UTI data. The following sections review the most relevant methods for analysing UTI data starting with manual video analysis, which is quite labour-intensive, but can be viewed as the baseline method. Next, tongue contour extraction is perhaps the most widely used way of analysing UTI data, but also suffers from efficiency issues. The last section introduces three dimension reduction methods with potential to be developed into a visualisation tool and to be used for automated onset detection.

2.4.1 Manual video analysis

Impressionistic, informal analysis and on-the-fly clinical interpretation can be very useful tools when working with UTI data; the former for checking data quality and as a preliminary step before more detailed qualitative analysis, and the latter in showing a therapist what their client actually does with their tongue (Preston et al. 2017, Cleland et al. 2019). Both require the user to be used to analysing and interpreting the data, because perceiving the anatomical structures in the images is difficult without training. Anecdotal evidence suggest that a month or two of working with the data improves a researcher's perception of the images making clear what previously was blurry and unclear.

Somewhat similarly to segmenting speech audio data, various types of speech video data – for example, ultrasound videos, lip and face videos of auditory speech, as well as videos of signed speech – can be segmented by going through frames, moving back and forth in time to pin point moments of change, and placing markers on these articulatory events. This, however, is at times quite difficult and time-consuming – partly because coarticulation blends articulatory gestures to each other, and partly because speakers will use different strategies for producing the same acoustic results.

2.4.2 Contour extraction

Anatomical features – mainly the tongue, but with appropriately recorded data also the palate (see above) may be extracted from the interpolated UTI data by various methods. This is usually done by fitting a spline to the tongue contour (or other feature to be extracted), which is usually referred to as 'splining' (Stone 2005). In the past, this has mainly been done manually, but in recent years there has been progress in automating the tongue contour extraction process (Fasel and Berry 2010, Lim et al. 2016, Xu et al. 2016). While the automated extraction methods are maturing, they still often require manual correction to be sufficiently accurate for analysis of fine detail in practical data analysis applications. This happens especially when the data contains noise and/or

artefacts (Csapó and Lulich 2015).

Figure 2.16 illustrates the process of splining a single ultrasound frame. In the left most panel (a) we have the interpolated ultrasound frame with some anatomical markers pointed out with arrows. The spline would be drawn on the frame following the bottom of the bright tongue surface echo like has been done in the middle panel (b). The final panel (c) presents the extracted contour outwith the context of the original frame.

Manual splining is slow work and whether the contour extraction is done manually or automatically it has two drawbacks: First, if the tongue contour lies parallel or almost parallel to the ultrasound beam, it produces no echo and can not be detected. This makes any tongue contour based on such data discontinuous or a matter of guess work. Second, contour extraction reduces the data in the image to a single contour disregarding all other data in the image. This means that information on the movements of any other structures besides the upper surface of the tongue are disregarded by spline analysis.

In the ultrasound image in Figure 2.16, we can see the mandible shadow to the right, the short tendon appearing as a brighter region to the left of the bottom of the mandible shadow, and other distinct regions inside the tongue. We can also see speckle above the tongue contour and while this is usually regarded as an artefact to be ignored, it is actually the product of indirect ultrasound echoes within the tongue (Stone 2005). And because it is produced within the tongue, changes in it reflect changes within the tongue even if they appear to be just speckle noise. After all the tongue is a muscle and muscle activation data is known to be available in the speckle noise in ultrasound data of muscles (See Koppenhaver et al. 2009, Vasseljen et al. 2009, and discussion in Section 2.3.6).

2.4.3 Dimension reduction methods

There are a number image analysis methods which do not rely on demarcation of structures in the images, and which can be applied to analyse change in UTI data. The most relevant ones to this thesis are Principal Component Analysis (PCA), optic flow, and methods based on Euclidean distance.

Figure 2.16: Extracting the tongue contour with a spline from a UTI frame. a) A UTI frame with arrows pointing to the top of the tongue surface near its tip (red, solid), tongue root (blue, long dashes), and the position of the ultrasound probe (green, short dashes), as well as some of the internal structure of the tongue visible, b) spline tracing of the tongue contour on the UTI frame (light green line), and c) the extracted spline (in black).

PCA and similar statistical dimension reduction methods have been applied to various types of articulatory image data such as cineradiography (X-rays), MRI and UTI (see for example Laprie et al. 2014, Engwall 2000a, Hueber et al. 2007). PCA is a general statistical method for finding the most salient components in data. As shown in the references, PCA can be useful in automatic recognition and modelling of articulatory gestures. However, the components that result from PCA are rarely intuitive for humans, and thus not the best tool in building a system of computer assisted articulatory data analysis. PCA also requires that it be set up for the specific data set to be analysed: changing image size or resolution require recalculating the principal components.

Optic flow (Horn and Schunck 1981, Raudies 2013) is a general image sequence analysis method which can be used to analyse direction and magnitude of motion evident in the sequence. An advantage of using optic flow is that it does not need any description of what the moving structures are. It simply finds areas of pixels that seem to have moved across the image, based on local correlations or similar metrics between subregions of each consecutive image. Optic flow has been used in analysing regular videos of speakers (Barbosa et al. 2008) as well as ultrasound videos (Moisik 2010, Bird et al. 2010). The main drawbacks are that optic flow algorithms can be computationally heavy and that the results of optic flow analysis are not necessarily straight forward to interpret. The latter problem can be eased by defining regions of interest corresponding to certain anatomical features and limiting the analysis to these regions (Barbosa et al. 2008).

Euclidean distance can be used to calculate the general amount of change in UTI or similar data. This approach has been used previously by McMillan and Corley (2010), Raeesy et al. (2011). It is an attractive method because it is simple to implement – Euclidean distance is the sum of squared differences, in this case differences at each pixel – and does not require adapting for different resolutions or other image parameters. The raw analysis result – Pixel Difference (Pixel Difference (PD)) from here onwards – is a time dependent curve describing the amount of change from one frame to the next.

2.4.4 Summary

Part of the data in this thesis has been segmented by manually analysing the ultrasound videos. This data set is used in testing the performance of the manual and automated methods developed in the thesis.

The method development of this thesis is based on the Euclidean distance (McMillan and Corley 2010, Raeesy et al. 2011). Compared to splining it has the advantage of using all of the data available in the ultrasound images and having an easy way forward for developing visual representations of the change present in a video sequence. Compared to the other dimension reduction methods, it has the advantage of being a simple measure of the difference between consecutive video frames and thus well suited for finding transition points such as articulatory onset. The development and testing of tools for calculating and analysing PD is described in detail in Chapter 3.

2.5 Research goals, research questions, and structure of empirical activity

This thesis has two main goals. First, it will analyse pre-speech tongue movements and relate their timing to the timing of the whole utterance. Second, to facilitate reaching the first goal and as a contribution to future research, it will be necessary to develop methods for quantitative, efficient and replicable analysis of UTI data.

To achieve the first goal, the thesis answers the research questions listed in the next two sections. It will be achieved by analysing data from three experiments, as described below. The focus, as we will see, is on the latter two experiments. The first experiment will primarily contribute to methodological exploration, as detailed immediately after the research questions are set out.

Figure 2.17 shows a model of the speech preparation timeline relating to the reaction time experiments to be undertaken, showing the regions of interest for the research questions. The timeline is an adaptation of the model for execution

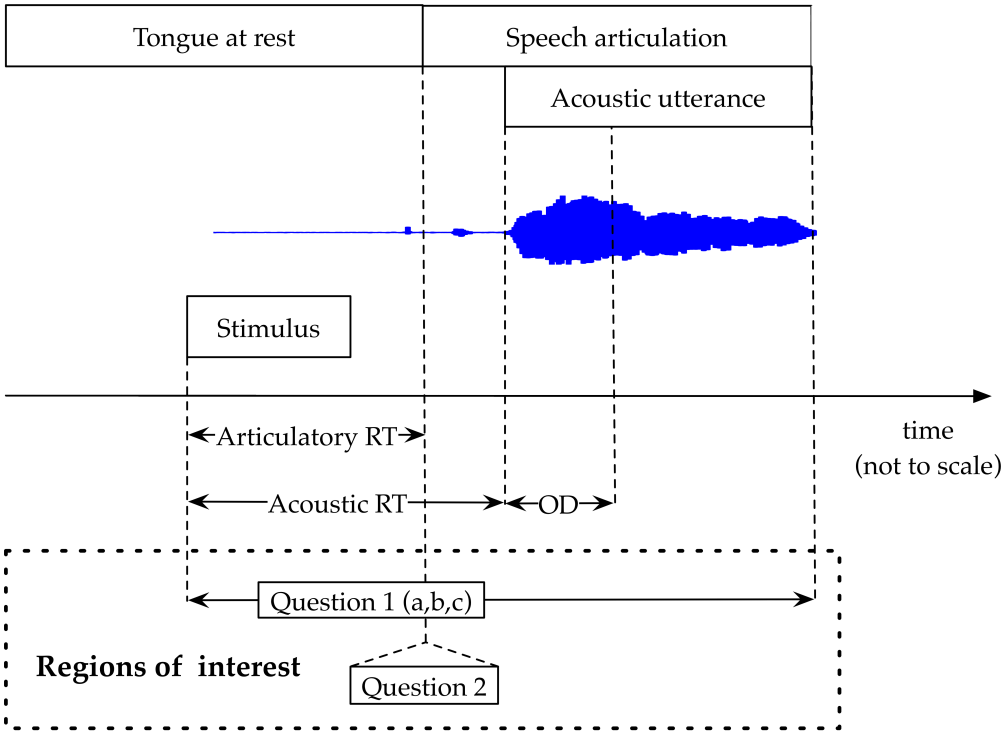

Figure 2.17: Regions of interest for the research questions in relation to the speech initiation timeline. OD stands for Onset consonant's acoustic Duration.

of prepared utterances by Sternberg et al. (1988) which was also used by Rastle et al. (2005) and discussed in Section 2.1.7.

2.5.1 Research Question 1: Timing of utterance onsets

What is the relative timing (and absolute reaction time in relation to the go-signal) of tongue movement initiation (or articulatory reaction time) and acoustic initiation (or acoustic reaction time) in different phonetic contexts in a speech reaction time task, following instructions used by Rastle et al. (2005)?

This a general question that will be answered by answering the more specific questions 1a, 1b, and 1c concerning respectively the articulatory reaction time, the acoustic reaction time and the AAI, which is the interval from articulatory initiation (reaction time) to acoustic initiation (reaction time).

Question 1a: Is articulatory reaction time affected by the acoustic duration of the onset consonant (OD) or by the acoustic duration of the utterance's rhyme?

Hypothesis 1a: Articulatory reaction time measured from the tongue is expected to depend on OD but not on the duration of the utterance's rhyme. The hypothesis is based on evidence from Mooshammer et al. (2012), but as discussed in Section 2.2, this may also be due to variation in syllable structure and a difference in the task. As there are no prior studies of the relation of articulatory reaction time and the duration of the utterance's rhyme, we have to choose the more conservative hypothesis.

The rhyme duration acts here as a proxy for the effect of speech rate and it will be defined as the duration of acoustic interval from the *end* of the onset consonant to the release of the final consonant in /CVC/ and /VC/ words. Chapter 5 gives more details on this.

Question 1b: Is acoustic reaction time affected by the acoustic duration of the onset consonant (OD) or by the acoustic duration of the utterance's rhyme?

Hypothesis 1b: Acoustic reaction time is expected to be inversely correlated with OD and positively correlated with rhyme duration. The first part of the hypothesis is based on the data reported by Rastle et al. (2005). The second part is based on the definition of a speech gesture in Articulatory Phonology as not just the constriction plateau but also inclusive of the onset and offset gestures (Browman and Goldstein 1988). Based on this it can be assumed that when an utterance is produced in a shorter time – and thus with a shorter rhyme duration – the articulatory onset gesture will also be produced in a shorter time. This in turn means that the acoustic onset will happen earlier, making the acoustic latency period shorten as the rhyme duration shortens.

Question 1c: Is the Articulatory to Acoustic onset Interval (AAI) affected by the acoustic duration of the onset consonant (OD) or by the acoustic duration of the utterance's rhyme?

Hypothesis 1c: The AAI is expected to correlate inversely with OD and positively with rhyme duration. Since we hypothesise that the acoustic reaction time will change as function of OD, we have to assume that the AAI will behave the same way. As for the rhyme duration, in terms of Articulatory Phonology, the AAI consists of the onset of the consonantal gesture and possibly part of the constriction plateau (depending on the sound being produced). It is thus hypothesised that the AAI will correlate positively with the time taken to produce the utterance and therefore with the rhyme duration.

2.5.2 Research Question 2: Difference between EMA and UTI

Do Ultrasound Tongue Imaging (UTI) and Electromagnetic Articulography (EMA) provide the same view of articulatory onset independent of the phonetic onset of an utterance?

This question is motivated by the difference between how vowel onset words pattern in results of Experiment 2 in Chapter 5 and in the results of Mooshammer et al. (2012).

Hypothesis 2: There are no previous direct comparisons between speech initiation data from UTI and EMA. However, since there is a difference in the results of Mooshammer et al. (2012) and those of Experiment 2, we hypothesise that there is a difference in how UTI and EMA see the articulatory onset. More specifically, we hypothesise that in UTI the vowel onset words will align with the inverse correlation pattern as if they had an onset consonant of 0 ms acoustic duration (OD = 0 ms), and that in EMA they will not align with the inverse correlation pattern because they will have a lengthened articulatory onset time.

Other possible explanations for the observed difference in the results include differences in the experimental paradigm (wait at rest vs. wait while vocalising), and differences in the anatomy and/or articulatory strategies of the participants.

2.5.3 Method development

Two change metrics that are based on the concept of Euclidean distance are developed in the next chapter. The first metric of these is called Pixel Difference (PD) and is loosely based on previous work by McMillan and Corley (2010), Raeesy et al. (2011). Its usefulness is explored in qualitative analysis of data from Experiment 1. PD is also used to implement a manual and an automated detection method for the articulatory onset.

The second metric is a novel refinement of PD. It is called Scanline Based Pixel Difference (SBPD) and it provides a metric for analysing localised change in ultrasound data. The localisation is based on the ultrasound fan and the use of raw data (Section 2.3.6 and the next chapter). Scanline Based Pixel Difference (SBPD) is further used to develop an improved automated onset detection method. The performance of the onset detection methods (both PD and SBPD based) is verified with cross correlation of their results with manual video analysis in Chapter 4.

2.5.4 Experiments

Experiment 1: This dataset is from a picture naming experiment which was recorded at Queen Margaret University in a separate project (Schaeffler et al. 2014) pre-dating this thesis project. The experiment task was picture naming of coloured versions of the Snodgrass standard set of pictures (Snodgrass and Vanderwart 1980, Rossion and Pourtois 2004). Data in the experiment was recorded with UTI and audio.

Data from picture naming experiments is characterised by relatively frequent false starts and hesitations. Furthermore, the target words are not matched in syllabic structure nor designed to provide systematic variation of the identity of the onset segment. These qualities limit the use of this data set to exploratory purposes in the present context of this thesis.

Since this dataset had already been recorded before the beginning of this project, it was used in early development and testing of the UTI analysis tools. The analysis of this data set presented in Chapter 4 is an exploration of the way PD represents change as a function of time and is mainly qualitative.

Experiment 2: This experiment was designed to answer Research Question 1 and its sub-questions 1a, 1b, and 1c. It is an adaptation of the delayed naming experiment of Rastle et al. (2005) with some necessary changes in the materials. Data from the experiment was recorded with UTI and audio. In the experiment lexical English /(C)(C)(C)VC/ words were produced by Scottish English speakers.

The hypothesis concerning Questions 1a, 1b, and 1c are all confirmed by fitting statistical models to the data of this experiment. Question 1 is answered by drawing together these results. The discrepancy between the results on /VC/ words in this experiment and the results of Mooshammer et al. (2012) is discussed. The articulatory onset locations (on the front-back continuum provided by the ultrasound fan data) are analysed qualitatively, but they do not point to a clear explanation of the discrepancy. This motivates Research Question 2 and Experiment 3.

Experiment 3: This experiment is a case study designed to directly compare UTI and EMA by removing the confound of speaker specific variation in articulatory strategies, anatomy, and cognitive factors. The experiment replicates Experiment 2 with UTI and EMA recorded at two different occasions. Because the only participant (the author) is a native Finnish speaker, the materials were phonotactically legal Finnish /(C)V/ syllables. Statistical analysis of the results shows that UTI and EMA are otherwise in agreement, except for a level difference between reaction times – both articulatory and acoustic – between the two modalities.

The next four chapters form the empirical part of this thesis. The next chapter covers the method development and the three following chapters each

cover one experiment. Answers to the research questions presented above are discussed in Chapter 7 before a more general discussion of the findings of this thesis.

Chapter 3

Method development: Pixel difference

To a casual observer, ultrasound videos of the tongue contain a lot of visual noise in addition to useful information about the location, shape, and movement of the tongue. Even in still images (Figure 2.16a) this can be seen as the blurriness characteristic of ultrasound images. The usual way of analysing Ultrasound Tongue Imaging (UTI) data is to take point measures and manual or half-manual tongue tracings based on the acoustic segmentation. As discussed in Section 2.4, such annotation is very laborious and time-consuming and a great deal of research effort is put into methods to automatically detect regions and boundaries (for example Xu et al. 2016, Lim et al. 2016). Furthermore, this approach discards data from the interior of the tongue (Figure 2.16b&c). Ultrasound data contains information about the activation and relaxation of muscle fibres – data which is crucial in detecting movement onset as early as possible (Koppenhaver et al. 2009, Vasseljen et al. 2009).

This chapter describes two change metrics and two movement onset detection methods based on the change metrics. One change metric is an adaptation of existing methods, and the other is a novel development. The change metrics aim to take into account the muscle activation data. The first of these is the pixel difference, which looks at the Euclidean distance between two images such as

ultrasound frames (McMillan and Corley 2010). These are interpreted to be N dimensional vectors, with each of the N pixels presenting a dimension. This difference is then calculated for each pair of images in a given video sequence resulting in a contour describing the amount of change over the whole sequence as a function of time.

McMillan and Corley (2010), Drake et al. (2013a;b) have used pixel difference in analysing downsampled UTI videos. Another application of the same basic method is reported by Raeesy et al. (2011) who use it in analysing Magnetic Resonance Imaging (MRI) data and combine it with acoustic analysis to show a correlation in the dynamics of the speech signal and the dynamics of the pixel difference. Lammert et al. (2013) select a region of interest automatically in vocal tract MRI data. They choose the pixel (or more correctly voxel in MRI) with the greatest change in a cine-MRI sequence and centre the region of interest on that pixel and then compute the pixel difference within that region of interest.

In an effort to provide fast tools, which are relatively light in terms of human workload, this thesis develops two pixel difference methods. The first one is called basic pixel difference algorithm. In contrast with McMillan and Corley (2010), Drake et al. (2013a;b), it utilises raw uninterpolated ultrasound frames with no downsampling (see next section for details and explanation). It is shown to be useful in determining articulatory onsets and in classifying tokens into categories based on the amount of hesitation evident. The algorithm is described in detail in the next section, followed by sections on selecting the best time step or frame comparison distance for Pixel Difference (PD) (Section 3.2) and on manual and automated onset detection based on PD (Section 3.3).

The second pixel difference method is called scanline-based pixel difference (Scanline Based Pixel Difference (SBPD)). It is a novel development based on the basic pixel difference, but with a refinement that gives access to movement location data. It is described in Section 3.4. Using it in identifying the point of first movement on the front-of-tongue – back-of-tongue dimension is described in Section 3.5. It is also used to develop the final articulatory onset detection method which is described and assessed in Section 3.6.

The final section of this chapter summarises the method development results, gives the reasoning for choosing the methods used for analysing ultrasound data in later chapters, and discusses using the methods in other projects.

3.1 Basic Pixel Difference (PD) algorithm

This version of PD was developed as the first step of method development in this thesis. The ultimate goal – which is mostly beyond the scope of this thesis – is to provide fully automated tools for analysing articulatory data. PD is a promising metric for measuring among movement in a video sequence, and provides an essential step towards the intermediate goal of developing an automated technique for detecting articulatory onset.

The version of pixel difference algorithm used in this thesis utilises raw ultrasound data. In this context raw data or raw ultrasound data refers to

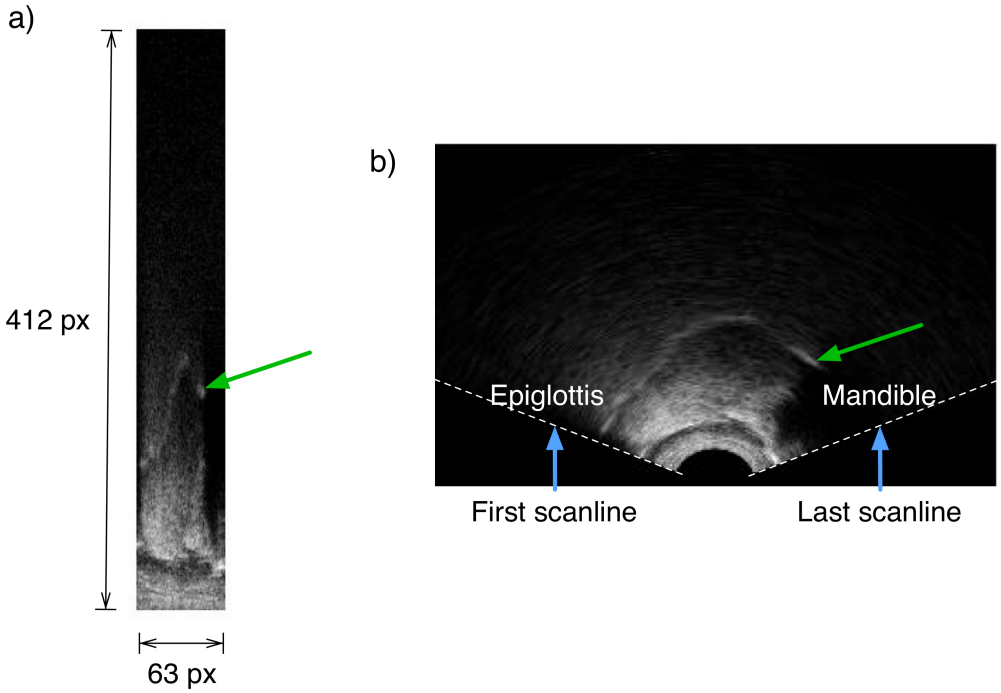

Figure 3.1: a) Raw (uninterpolated) version of an ultrasound frame from the data of Experiment 2. b) Interpolated version of the same ultrasound frame. The green arrow points to tongue tip in both a) and b).

uninterpolated probe return data as illustrated in Figure 3.1 and explained earlier in Section 2.3.6. In a nutshell, it is data formed of scanlines where each scanline corresponds to the echo data recorded by the ultrasound probe in response to an ultrasonic ping it sent into the imaged tissues. Crucially for the algorithms presented in this and following sections, each scanline in each frame is a vector of discrete values ranging between 0 and 255 – as is, consequently, the whole frame.

To calculate the PD between two UTI frames we interpret each raw frame as a $N = n_x \times n_y$ dimensional vector. The PD $d1$ between consecutive UTI frames is then defined as the Euclidean distance between the two frames im_k and im_{k+1} with indices i and j iterating over the pixels in x and y direction ($im_k(i, j)$ denotes the pixel in frame k, at row i and column j):

$$d1(k) = ||im_k - im_{k+1}|| = \sqrt{\sum_{i=1}^{n_x} \sum_{j=1}^{n_y} (im_k(i, j) - im_{k+1}(i, j))^2} \tag{3.1}$$

for $k = \{1, 2, \ldots n_{frames} - 1\}$.

The calculations and the resulting PD contour as a function of time are demonstrated in Figure 3.2. The figure shows a mock up sequence of ultrasound frames consisting of 4x3 pixels with values between 0 and 255 – as in real raw ultrasound data. In the simulated sequence, the first two frames have only a change in noise between them. Between the next two – numbers 2 and 3 – the tongue contour moves a bit resulting in a significant rise in PD. There is even more movement and hence more change between frames 3 and 4 resulting in the largest PD value in this example. There is a change in noise between the final two frames – like between each of these frames, but more importantly there is a blurring of the left most pixel of the simulated tongue contour, which results in a higher PD value then between the first two frames.

The difference can be calculated as readily for images further removed

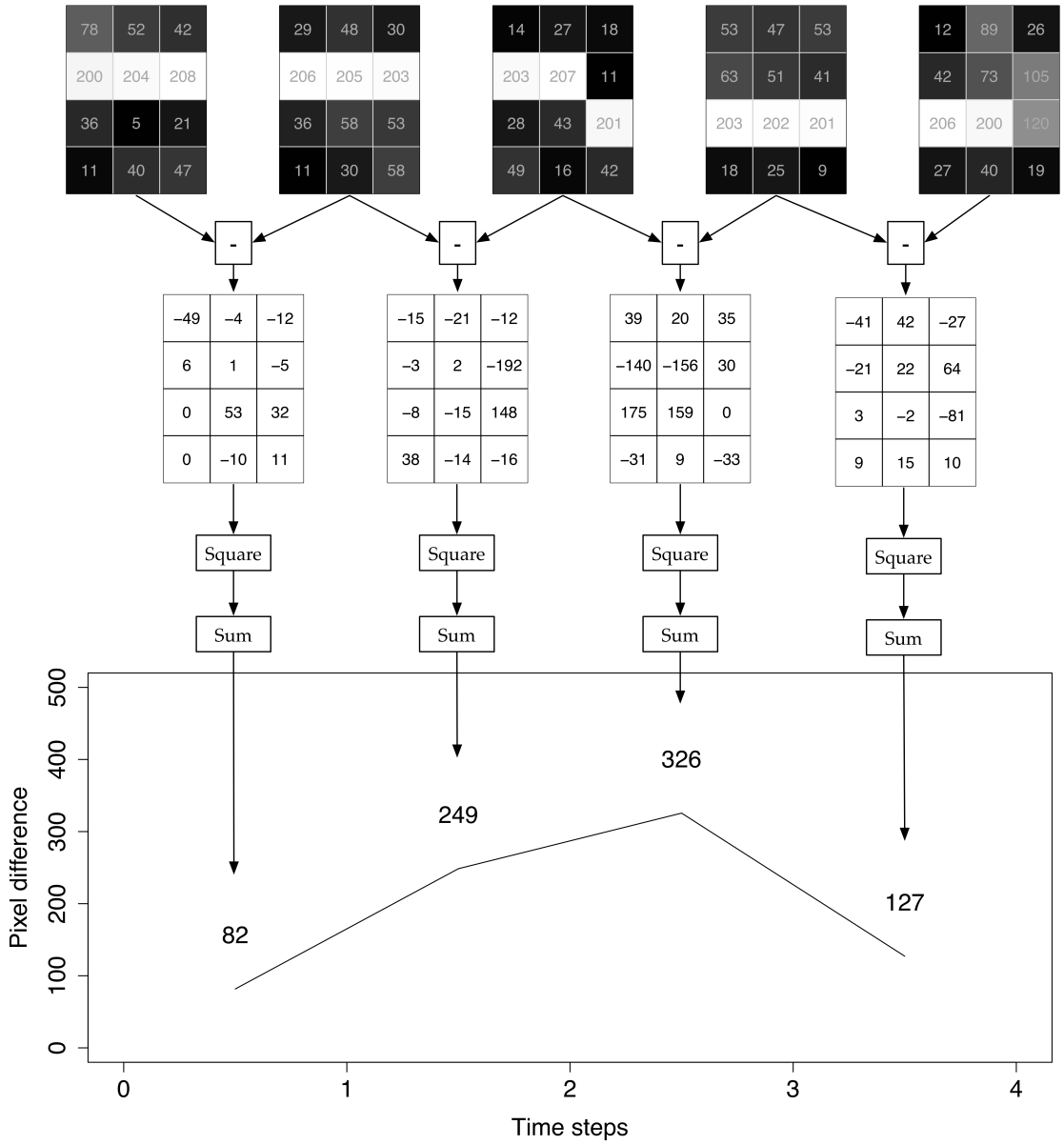

Figure 3.2: Calculating PD with $d1$ and the resulting PD contour demonstrated with a series of mock up frames consisting of 4x3 pixels and with simulated noise and a lighter band of pixels across each of the frames to provide a simulated tongue contour. To calculate the PD, we first calculate the change or difference between two frames for each pixel. After squaring these individual differences, we take the sum over the change matrix to get the total pixel difference between two frames.

(step n instead of step 1) and are defined as

$$dL(k) = ||im_k - im_{k+L}|| \tag{3.2}$$

for $k = \{1, 2, \ldots n_{frames} - L\}$. The time stamp $t_{dL(k)}$ corresponding to a single difference value $dL(k)$ is defined as the average of the time stamps of the corresponding UTI frames:

$$t_{dL(k)} = \frac{1}{2}(t_{im_k} + t_{im_{k+L}}), \tag{3.3}$$

where the time stamp of image k is the time its acquisition ends. In effect this means that the larger the value of L, the less localised in time the metric is.

Figure 3.3 illustrates how calculating $d1$ and $d3$ differ in practice for a mock-up sequence of ten frames consisting of 3x2 pixels. As can be seen, for a sequence of 10 frames, $d1$ will produce 10-1=9 PD values and $d3$ will produce 10-3=7 PD values. To relate these values to time we take the average of the acquisition time stamps of the frames used to calculate a given PD value. Thus, the time stamp for the first $d1$ value in the example would be an average of the time stamps of frames 1 and 2, and correspondingly for the first $d3$ value an average of the time stamps of frames 1 and 4.

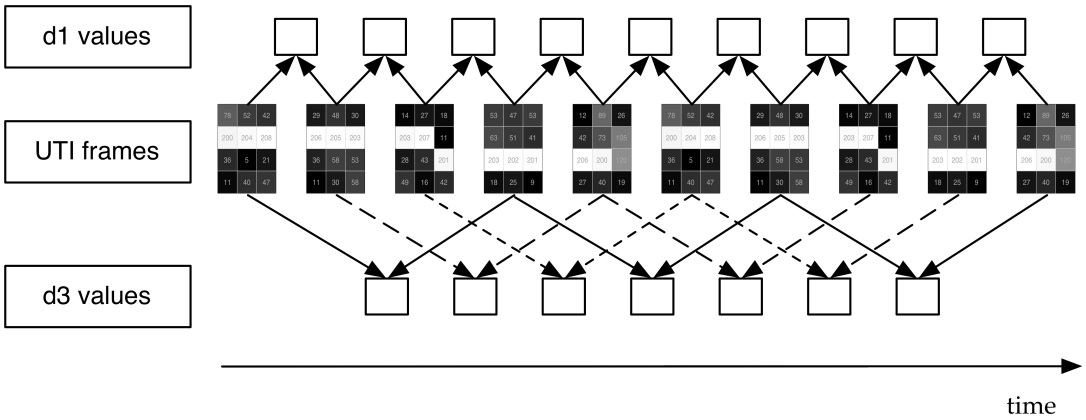

Figure 3.3: Relationship of differences $d1$ and $d3$ to UTI frames in a sequence over time.

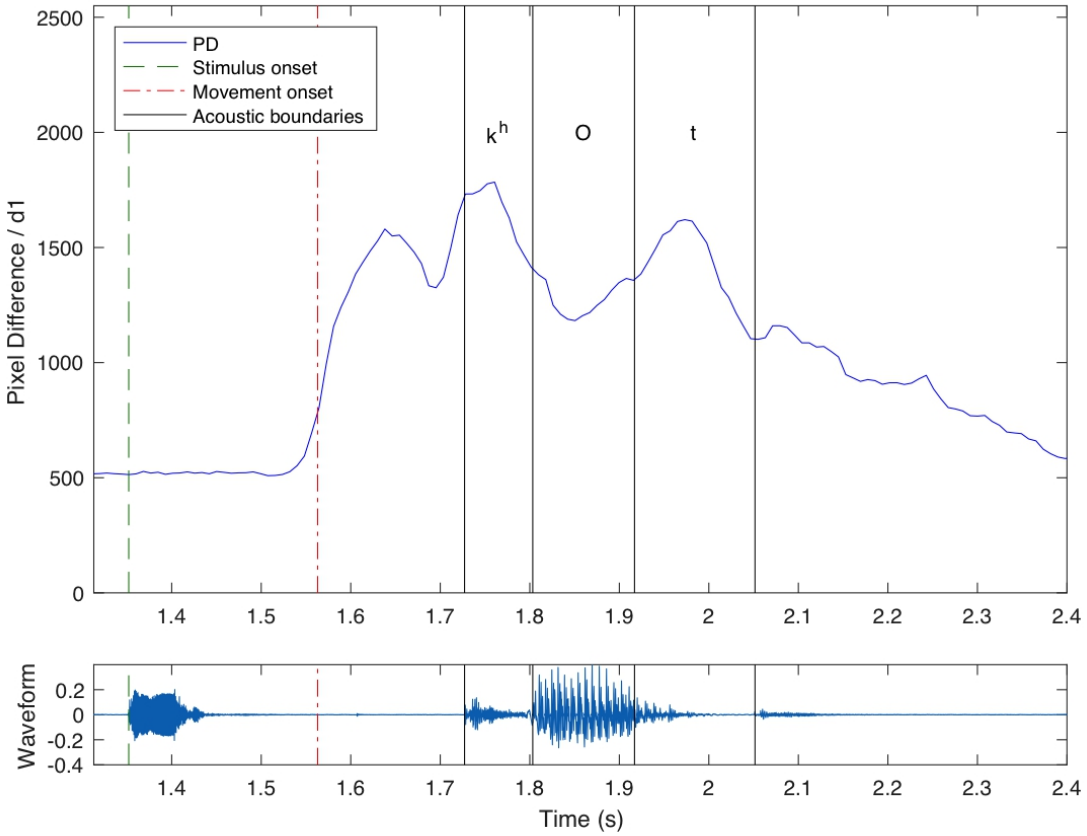

Figure 3.4: Example of a pixel difference contour and the corresponding acoustic waveform: Participant P1 of Experiment 2 reading the word 'caught'. Stimulus onset marks the go signal (a 50 ms long 1 kHz beep visible at the beginning of the waveform), movement onset was manually labelled by examining the UTI video, and acoustic boundaries were determined in Praat (Boersma and Weenink 2010).

Figure 3.4 shows an example of PD applied to real data with corresponding acoustic segmentation. The example is from Experiment 2. In it participant P1 reads the word 'caught' in a delayed naming context. The PD contour in the figure is typical of a steady production without any hesitations present. The PD starts at an almost steady level corresponding to no significant movement of the tongue. It is important to notice that even though there is no movement of the tongue, PD is not zero but rather at noise floor. The value of the noise floor depends on the scanner settings and the physiology of the participant. Being

essentially a product of a random process, the noise floor fluctuates a bit from one time instant to the next.

Further typical features of the PD curves are evident in the example. In clear examples like this one, manually labelled articulatory onset falls on the lower slope – but practically never on the bottom – of the rise to the initial peak. This means that PD registers movement onset before human annotators. Or more precisely this change metric registers change before human annotators detect movement.

Another feature is the general relationship of peaks and valleys to the acoustic boundaries. Here the first peak at about $t = 1.65$ s corresponds to the maximum change during the closing gesture of [k], while the second peak at about $t = 1.75$ s corresponds to the maximum change during the release gesture. Similarly, [t] has been acoustically labelled as the voiceless closure between approximately 1.91 s and 2.05 s and corresponds to a peak during the closing gesture with a following PD peak immediately after 2.05 s corresponding to the release gesture and its associated acoustic release burst visible in the waveform. Unlike the consonants, the acoustic vowel [o] corresponds to a valley around 1.85 s in PD.

The example in Figure 3.4 is on purpose a clear one. As we will see in the results of Experiment 1, PD is not always this clear. As a potential way for clarifying the location of articulatory onset in less clear utterances, we will next consider the issue of selecting the time step to be used in PD analysis.

3.2 Selection of the best time step for PD

From the definition of the dL metrics, it is clear that $d1$ has the best time resolution or localisation in time. This is a direct result of comparing adjacent frames. Taking a longer time step means that we are comparing images that are further apart in time, and thus the time window that we are observing spreads. For ultrasound operating at approximately 120 fps (like in Experiment 2 and 3), this means that the window lengths corresponding to $d1$, $d3$, and $d5$ are

respectively 8.33 ms, 25 ms, and 41.67 ms.

Because this time window moves by steps that are equal to the time steps between frames, it would be wrong to think of the window length directly as time resolution: Data points of any dL metric are set apart by the same time step. However, the greater the length of the time window or comparison step, the less localised the metric is in time.

To understand how this affects analysis of a UTI sequence, consider a case where we are analysing an ultrasound sequence with the above frame rate of 120 fps and there is a sudden change in change rate at time t_{onset}. If we use $d1$ for analysis this change will be first detectable in the comparison between the frame at t_{onset} and the immediately preceding frame because this is the first comparison involving the frame with the sudden change. In contrast, if we use $d5$, the change will be detectable four frames earlier, thus, effectively spreading the representation of the change over a longer time.

If the question were of only optimising time resolution of the method, then it would be clear that we should prefer the shortest time step. However, as we can see in Figure 3.4, PD contains non-negligible levels of noise. So, we should consider sacrificing time resolution, if we can improve the signal-to-noise ratio – and perhaps make visible changes that are lost in the noise when using $d1$ – by doing so.

To make this decision two example cases in which we would expect a speaker to be stationary or nearly stationary were analysed. In the first case, the speaker stayed motionless and silent before speaking out. In the second case, the speaker produced a prolonged vowel. Figure 3.5 shows the PD curves for the first case. The two panels in the figure show that something – three peaks at even intervals of about 0.45 seconds – does appear in the metrics with a longer step – something that is not visible in the $d1$ curve. Looking only at the upper panel it is already evident that employing a longer time step does not alter the shape of the curve significantly. This is confirmed by the plot in the lower panel where the metrics have been normalised by scaling each with their own maximum value: in the interval of speech articulation – the highest peaks

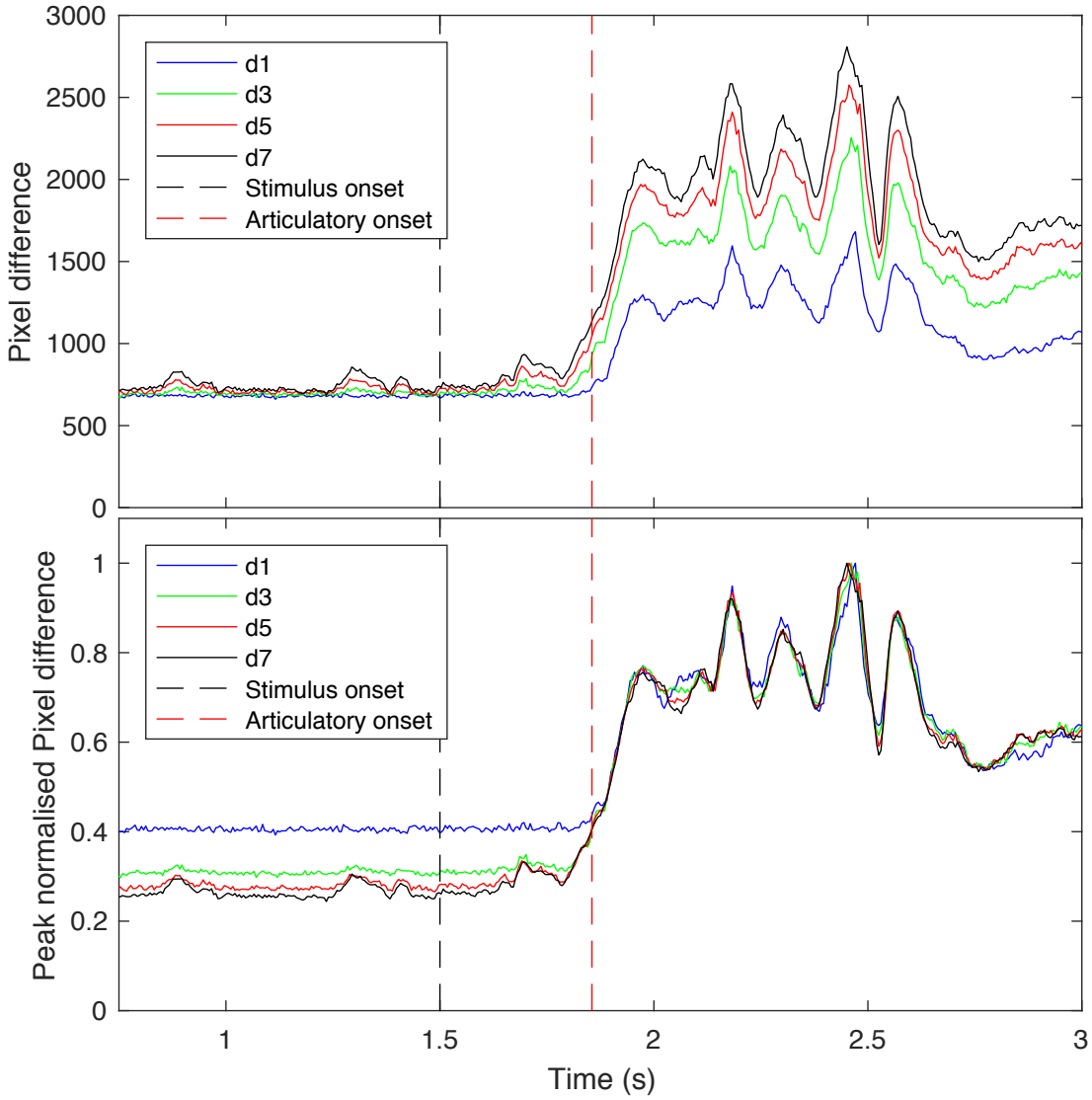

Figure 3.5: Relating video annotation to the PD contour: Participant E1 of Experiment 1 naming a picture of a snowman. Stimulus onset marks the appearance of the picture on the computer screen, and articulatory onset was manually labelled by examining the UTI video and annotating the first point in time where the labeller perceived tongue contour movement.

– the different metrics lie practically on top of each other. As the bottom panel demonstrates the relative noise floor does actually drop for the metrics with a longer time step in relation to $d1$.

Figure 3.6 shows the PD curve of the second case and corresponding waveform. This recording is of a 10-second-long Finnish /a/ produced by the author. Like in the previous figure the pixel difference contours show a periodic structure. This time it appears during the 10 seconds of holding the same articulatory position. This means that the peaks are not associated with linguistic change since they appear in a period where there a priori is none. Furthermore, the period of these peaks and those in the previous figure is in the range of a regu-

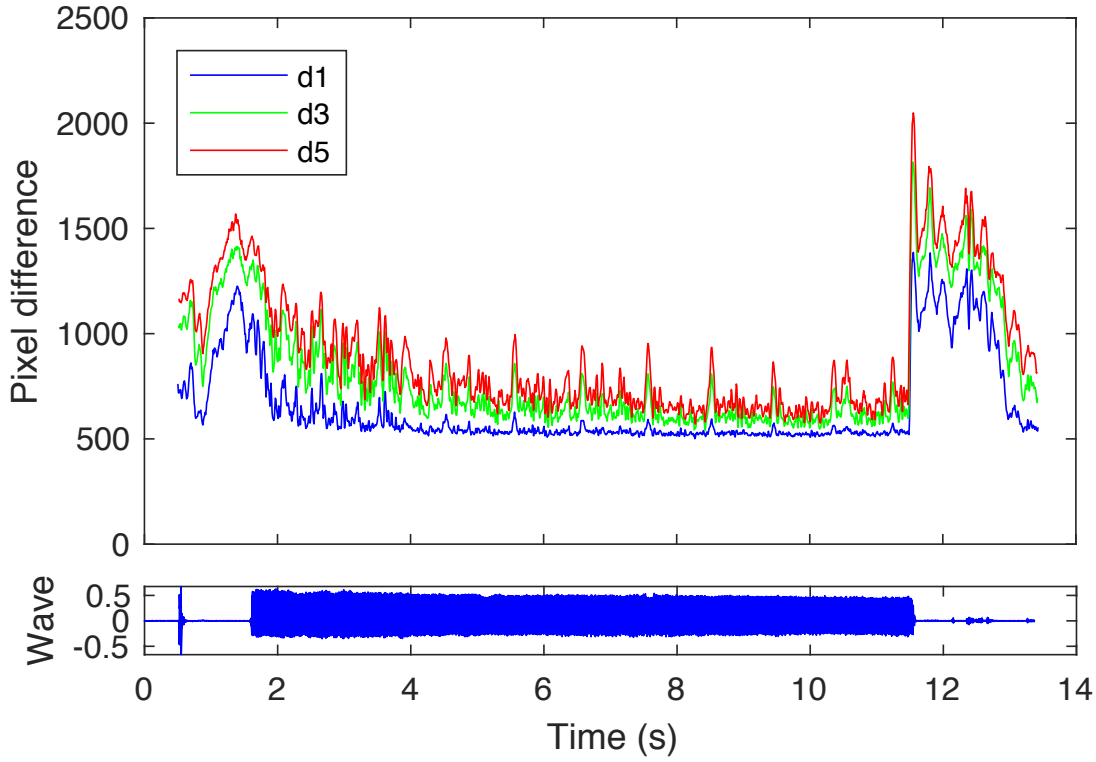

Figure 3.6: A prolonged /a/ produced by the author. The steady vowel lasts about 10 seconds. After the noisier first two seconds of the stationary articulation (time stamps between 2 and 4 seconds on the x-axis) the $d1$ contour flattens except for a steadily repeating peaks that repeats 8 times over about 7.5 seconds. The same peaks are more pronounced in $d3$ and $d5$.

lar heartbeat. Examination of the ultrasound video corresponding to Figure 3.6 revealed no significant movement of the tongue contour during the steady 10 second articulation. It did, however reveal movement in a small area inside the tongue. That area is a pulsing blood vessel, thus, verifying that the peaks are a result of the author's pulse within the tongue.

Since there is no new linguistic information to be gained by using a longer time step and doing so would degrade the time resolution of the metric, $d1$ is used exclusive in analysing ultrasound data in this thesis.

3.3 Manual and automated onset detection based on PD

As discussed earlier in Chapter 1 and Section 2.4.1, analysing and annotating articulatory data can be very time-consuming. Moreover, principled metrics that would provide useful alternate views of the data – like the spectrogram does for sound – are not freely available for ultrasound data.

To address these needs an annotation tool for viewing and manual labelling of PD contours has been implemented in Matlab. From pre-computed PD data and pre-processed audio data, the tool produces a graph showing the PD curve as a function of time with the timing of the go-signal and acoustic onset (if available) marked on the graph. The annotation tool includes an automated onset detector, which uses dynamic time warping to identify articulatory onsets based on PD.

The basic principle of dynamic time warping is illustrated in Figure 3.7. It works by matching the behaviour of two time series signals as closely as possible by locally compressing and stretching the time dimension of each signal until a best fit is found (See Müller 2007, for more details). In this thesis, an exponential test function was used to match the rising slope of articulatory onset in the PD contours.

The annotation tool can be used to speedily label articulatory onsets on PD curves by running the automated onset detector first and then manually

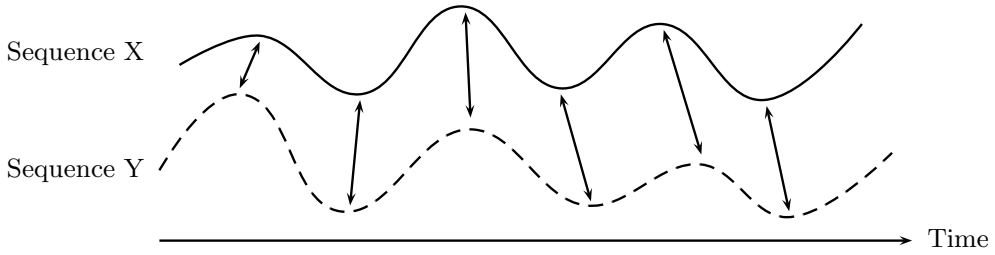

Figure 3.7: The principle of matching two signals with dynamic time warping (Figure from Müller 2007).

correcting the results before moving to the next token. The code for the tool is available in Appendix B. In this thesis the annotation tool is used for labelling articulatory onset on PD contours and for categorisation of the general types of PD contours. The contour types are discussed further in Chapter 4.

3.3.1 Comparing articulatory annotation methods

Three annotation methods were tested using a test set of delayed naming tokens from Experiment 2. The set consisted of 333 /VC/ and /CVC/ tokens from the data of the first session recorded with participant P1.

First, the test set was annotated by the author by manually searching through the ultrasound videos for the movement onset in Articulate Assistant Advanced (AAA). The user interface in AAA gives the opportunity to play the ultrasound data as a video or to page forwards and backwards through the data one frame at a time. At first when getting to know the data the first option was used, while the second option was used for the actual annotation.

Second, it was automatically annotated with the annotation tool without any human intervention. The automated annotation utilises only the PD data as described above.

Third, the set was interactively annotated by three phoneticians with the annotation tool. While all three judges were experienced with ultrasound, only Judge 1 (the author) had any greater experience with PD. Go-signal start and end, and manually corrected acoustic onset of speech (see Chapter 5) were

displayed on the PD curve to guide the annotators work. Judges 2 and 3 were instructed in using the annotation tool and shown some examples of PD curves and where to mark the articulatory onset on them.

3.3.2 Speed of annotation and data loss

Annotation sessions were not formally timed because in terms of time taken to annotate the test data set, the three different methods have a clear order: manual video annotation is slowest, and fully automatic annotation is fastest. Video annotation of the test set took 180 minutes and was done in three batches of 60 minutes. The task is more strenuous than audio annotation and breaks are needed to keep the annotator from either slowing down or becoming less reliable. For both the automatic annotation and manual annotation with the analysis tool a necessary preparation step is to calculate PD for all of the tokens to be analysed. For 333 tokens this takes about 30 minutes of unsupervised processing time (i.e., it can be run in the background while working on other tasks). Automatic annotation of the test set runs in negligible time. Each of the phoneticians annotated the whole data set with the analysis tool in about 30 minutes making manual PD annotation six times as fast as manual video annotation.

The PD method picks up changes in the state of muscle fibres inside the tongue (Koppenhaver et al. 2009, Vasseljen et al. 2009). These changes precede the movement onset of the tongue surface, but would not be as early the onset of Electromyography (EMG), because Electromyography (EMG) detects the change in electric potential caused by neural activation, whereas PD detects the movement of caused by that activation, which naturally starts slightly later.

Since we are exploring the workings of these methods, we will select a conservative lower bound as the exclusion threshold for SBPD onsets when removing outlier tokens. Using a conservative threshold means that we will remove only the most extreme outliers from analysis. Our threshold is based on an estimate which Chiu and Gick (2014) calculated from several sources in the literature. Chiu and Gick calculated the estimate as part of a study of STARTLE

type data. In the STARTLE paradigm, in a subset of the trials the participant is startled by a very loud (> 120 dB) go signal. This produces reaction times which are faster than regular ones (Chiu and Gick 2014).

Adapting Chiu and Gick (2014) (see also Section 2.1.4), we calculate the sum of the lower bounds of signal latencies along the neural path from the onset of an acoustic stimulus to cortical response (10 ms), the latencies of the cortical processes involved (total 7 ms), and the onset latency of a motor response defined as EMG onset (11 ms). The result is 28 ms and will be used as the exclusion threshold in throughout this thesis. We have omitted the final step of motor time (defined as EMG onset to movement onset and lasting 30 ms) to produce a less strict threshold.

Table 3.1 shows the number of tokens lost in each annotation method and in the annotations of each judge. In the manual video analysis, tokens were marked for exclusion from analysis by marking the articulatory onset before the go-signal. Since these are not distinguishable from tokens where the participant started moving before the go-signal, excluded tokens are not differentiated from tokens where the movement starts too early for manual video analysis. In manual annotation of PD the judges had the option of marking the token as missing data, if they felt that the token either was not analysable – that is, did not have a clear articulatory onset – or if the articulation onset preceded the go-signal. Because of these confounds of early starts and unclear starts in the annotation protocols, comparisons of lost data should be based on the total number of excluded tokens.

Table 3.1: Number of tokens excluded from the basic test data set for the three articulation onset annotation methods and for three different judges. Judge 1 annotated the data on two separate days (marked as Day 1 and Day 2).

	Manual	PD	Manual annotation			
	Video	Auto	Day 1	Day 2	Judge 2	Judge 3
Marked for exclusion	*n/a*	0	52	88	7	31
Annotation < 28ms	10	9	12	1	0	0
Total excluded	10	9	64	89	7	31

Manual video analysis, the fully automatic method and Judge 2, show similar levels of missing data. Judge 3 shows a clearly higher level of excluded data and Judge 1 has the highest levels of all on both days. Properly explaining these differences would be a matter of further experimentation, but it is possible that the higher exclusion rate of Judge 1 was due to his greater familiarity with the PD data.

For analysing agreement between the annotation methods in the following section, only tokens with complete data (no missing values from any method or annotator) were used. Combining the data from all annotators and annotation methods, there were 232 tokens in total with no missing annotation data and for which all annotators had marked reaction times which were ≥ 28 ms.

3.3.3 Correlations of the PD annotation methods

The results of the annotation experiment are presented in Figure 3.8, which shows scatter plots of the PD based annotation of each judge against other judges, the manual video annotation, and the automated method. We can see that overall there are clear correlations between all of the methods used. In the lower panels an identity line ($y = x$) lets us see that the PD based movement onset times are below those obtained from manual video annotation with very few exceptions which are mainly found as outliers of automated PD annotation.

To quantify this apparent level difference in the expected values of different annotation methods and judges, a simple ANOVA model was fitted to the data in R (R Core Team 2013), which was used for all of the statistical analysis in this chapter. The model was fitted with articulatory reaction time as the dependent variable and annotation method/annotator as the independent variable. The F-statistic was 60.502 on 5 and 1386 degrees of freedom resulting in a P-value $< 2.2 \times 10^{-16}$ (which is R's limit of precision on the computer used, meaning the P-value is ≈ 0). This means that at least one of the annotation methods produces results which differ statistically significantly from the results of the other methods.

Based on the result from ANOVA, Tukey's honest significance test was run

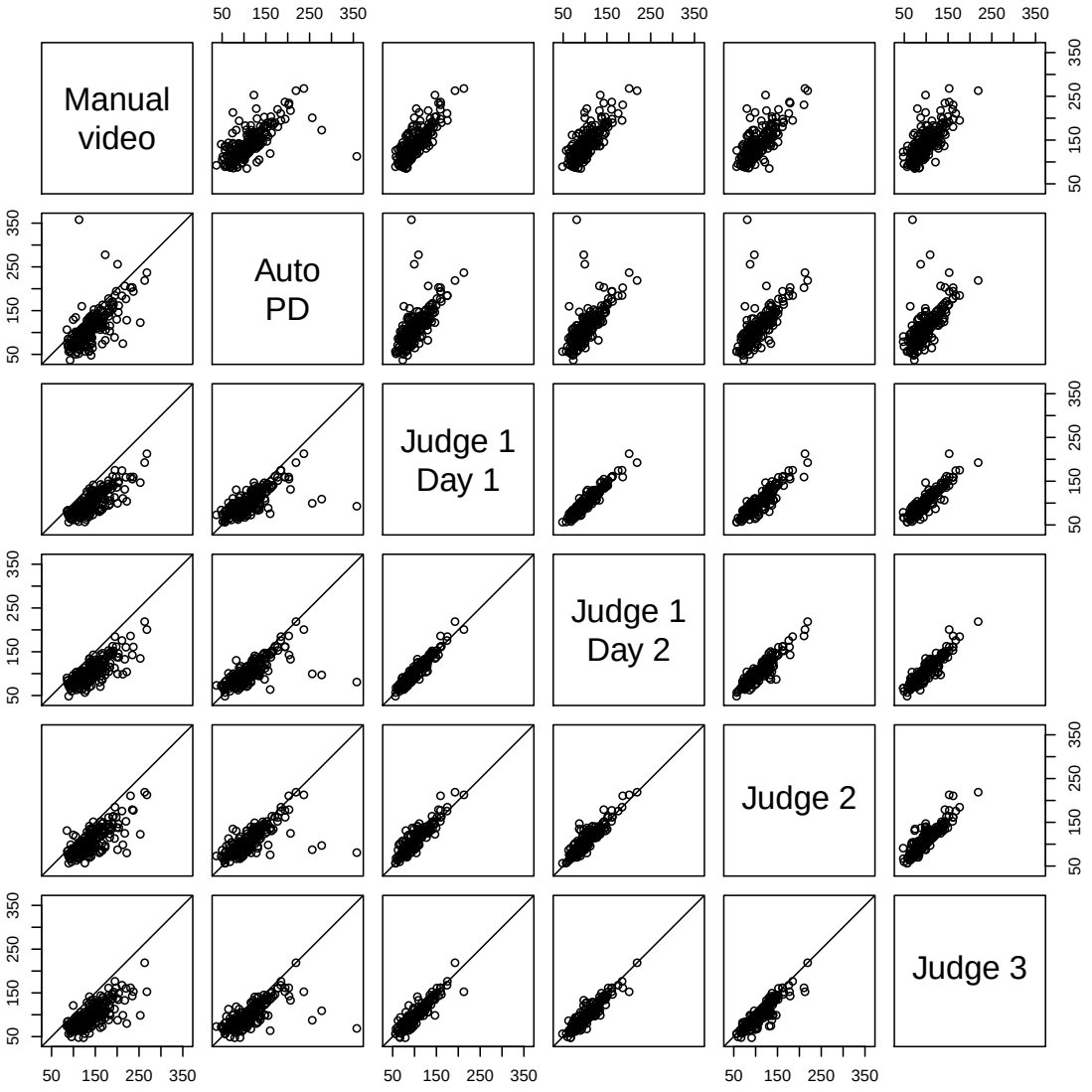

Figure 3.8: Scatter plots comparing manual video segmentation and automated articulatory onset detection with manual labelling of the PD curves by three human judges. Judge one (the author) labelled the data set two times on two separate days. The lower panels show also an identity line ($y = x$) to make it easier to see the general trends in the data. Scales are in milliseconds.

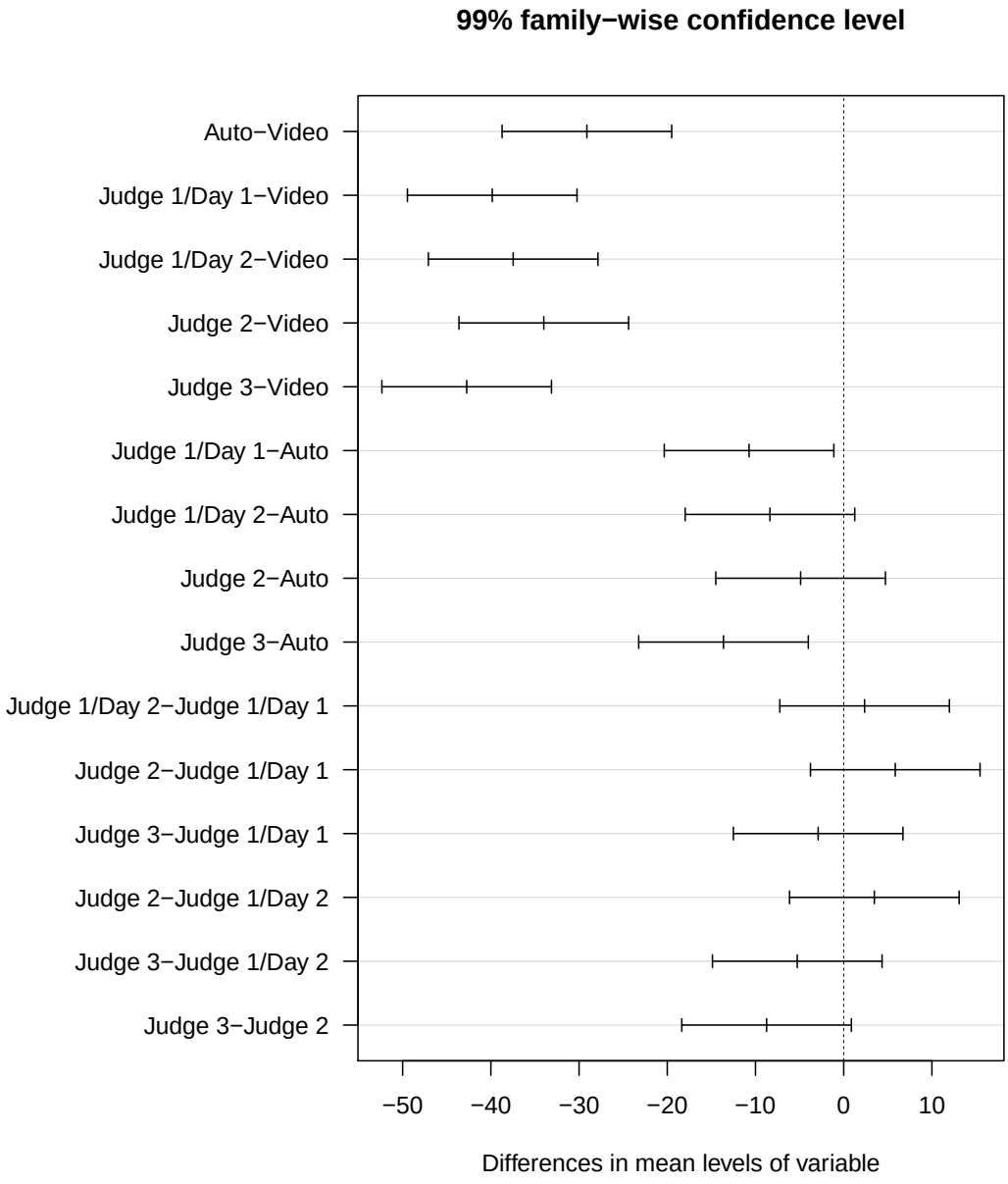

Figure 3.9: Confidence intervals ($P = 0.99$) for post-hoc pairwise comparisons of articulatory onset detection methods. Methods can be considered to have the same expected value if the confidence interval of their mean difference overlaps 0. Scale is in milliseconds.

Table 3.2: Pearson's correlation coefficients of manual ultrasound video labelling, manual labelling of the PD curves by three human judges, and the automated articulatory onset detection by dynamic time warping. Judge 1 (the author) labelled the data set two times (Days 1 and 2).

	Manual	Auto	Manual PD annotation			
	Video	PD	Day 1	Day 2	Judge 2	Judge 3
Video	1	0.657	0.817	0.789	0.750	0.736
Auto PD	0.657	1	0.690	0.676	0.660	0.650
Judge 1, Day 1	0.817	0.690	1	0.952	0.914	0.912
Judge 1, Day 2	0.789	0.676	0.952	1	0.920	0.920
Judge 2	0.750	0.660	0.914	0.920	1	0.895
Judge 3	0.736	0.650	0.912	0.920	0.895	1

with 99% confidence level. The results of the test are presented in Figure 3.9. We see that no other method produces results that could be considered equal (in the sense of expected values being equal) with the manual video annotation. Rather, manual video annotation produced the longest articulatory reaction times. The case of the fully automated measure against human annotators is less clear. We even find that the same annotator on two different days changes from non-equal to equal. Finally, all of the human annotators can be considered equal in their results with Judges 2 and 3 coming closest to having separate expected values. This leads us to conclude that the PD measures are on average more sensitive to small changes in the ultrasound data than manual video annotation.

Table 3.2 lists the Pearson's correlation coefficients between each judge, the onsets automatically identified by dynamic time warping, and the results from manually labelling ultrasound videos. Statistical testing shows that all correlations are statistically significant with all P-values $< 2.2\times10^{-16}$ when tested individually against the null hypothesis that $\rho = 0$ using a two-tailed t-test with 230 degrees of freedom. Examining the table of correlations, it is evident that the best agreement is between manual PD annotators, good agreement between manual video annotation and manual PD annotation, and only fair agreement between automatic PD annotation and any other method.

3.3.4 Summary

Achieving fair correlation with the video annotations is desirable, but not necessarily the best way to judge the usability of PD based onset detection, because PD is sensitive to tongue internal changes, which human video annotators tend to ignore.

In general, the judges show a good correlation among each other, but lower correlations with the results from video annotation and the numbers of excluded tokens are inconsistent especially within the modality of manual PD annotation. It is hardly surprising that there are individual differences, given that no second annotation of tokens was done as part of the experiment. We would, however, recommend doing so in the future, where more than one annotator annotates the same data.

The automated method has the lowest correlation values with all of the other methods. This is not surprising due to the outliers, which can be seen in Figure 3.8. They are, however, few in number and should not interfere with statistical analysis if proper outlier removal is performed.

One lucrative idea for improving the accuracy of automated onset detection, is to adopt the response locked approach used by van der Linden et al. (2014). Working EMG data they locate the activation spike that leads to articulatory movement by backtracking from the acoustic onset of speech to the first preceding spike.

A similar approach was tried for locating the PD onset. Given the nature of steady PD, which combines flat-line behaviour with small, sharp perturbations, the curves need to be smoothed before local minima can be identified. Unfortunately, it proved difficult to find a satisfactory method of smoothing the curves that would provide a reliable onset point when compared with the unsmoothed curves. Furthermore, this approach is also problematic in cases such as Figure 3.4 where the local minimum preceding the acoustic onset is actually not the movement onset, but the minimum between the closing gesture and release gesture of [k].

3.4 Scanline-based Pixel Difference (SBPD)

SBPD is an algorithm for computing PD for individual scanlines. It is a measure of the change present in the data from each scanline. It also makes it possible to calculate the articulatory onset for each individual scanline leading to a more robust automated measure of the overall articulatory onset as we will see in the next two sections. This development path was taken to explain the discrepancy between the results of Experiment 2 and those of Mooshammer et al. (2012).

With access to raw ultrasound data (Figure 3.1), the scanlines can be exploited to provide a local change metric instead of a global one like regular PD. In many ways, the principle and computations for SBPD are the same as those for PD set out in Section 3.1. There is, however, a computational difference between PD and SBPD. Instead of taking a sum over both scanlines and pixels along them we now only take the sum along each scanline and the result is a time dependent vector. Thus, for each scanline we have:

$$d1(sl,k) = ||im_k(sl,\dot{}) - im_{k+1}(sl,\dot{})|| = \sqrt{\sum_{j=1}^{n_y}(im_k(sl,j) - im_{k+1}(sl,j))^2}, \quad (3.4)$$

where sl stands for scanline and other notations are same as for PD.

When we repeat the calculation for each scanline we get, for each pair of ultrasound frames, a vector whose length is equal to the number N of scanlines in the data. Equation 3.5 gives SBPD in vector form with first element corresponding to the first scanline, second element to second scanline, and so on. We see that each element is calculated according to Equation 3.4.

$$\begin{pmatrix} d1(sl=1,k) \\ d1(sl=2,k)) \\ \vdots \\ d1(sl=N,k) \end{pmatrix} = \begin{pmatrix} ||im_k(1,\dot{}) - im_{k+1}(1,\dot{})|| \\ ||im_k(2,\dot{}) - im_{k+1}(2,\dot{})|| \\ \vdots \\ ||im_k(N,\dot{}) - im_{k+1}(N,\dot{})|| \end{pmatrix} \quad (3.5)$$

Figure 3.10 displays two examples of SBPD in the middle panels under the corresponding basic PD contours in the top panels. Acoustic boundaries

and the video-based articulatory onset have been marked on the graphs. The first example on the left shows a complete production of the word 'caught' also displayed in Figure 3.4. The second example on the right shows the beginning of the word 'sheet' being produced.

The scanlines are stacked first to last from top to bottom so that data that originates at the back of the tongue is displayed at the bottom of the graph while the top of the graph has data from the region around the front. The fact that the mandible is present in both videos accounts for the lighter shade on scanlines from 50 upwards because these correspond to the front of the image where the mandible shadow produces very little change from frame to frame.

Finer analysis of the features present in the graph during speech will require further work, but for the purposes of this thesis, it is important to note that in the second example it can be seen that movement does not begin in a uniform fashion across all of the scanlines. Instead, the region around scanlines 15-30 has the earliest change to darker shades indicating that movement onset happens in this region, which corresponds to a region towards the back of the tongue but not quite as far back as the image reaches. It should be also noted that care needs to be taken in interpreting this type of data: even when stabilised the probe moves during a recording session and maybe fixed in a very different position when the headset has been taken off between recording sessions. Thus, scanlines are not – nor are even groups of scanlines – an anatomically fixed region, but only indicative of the general region where the change happens.

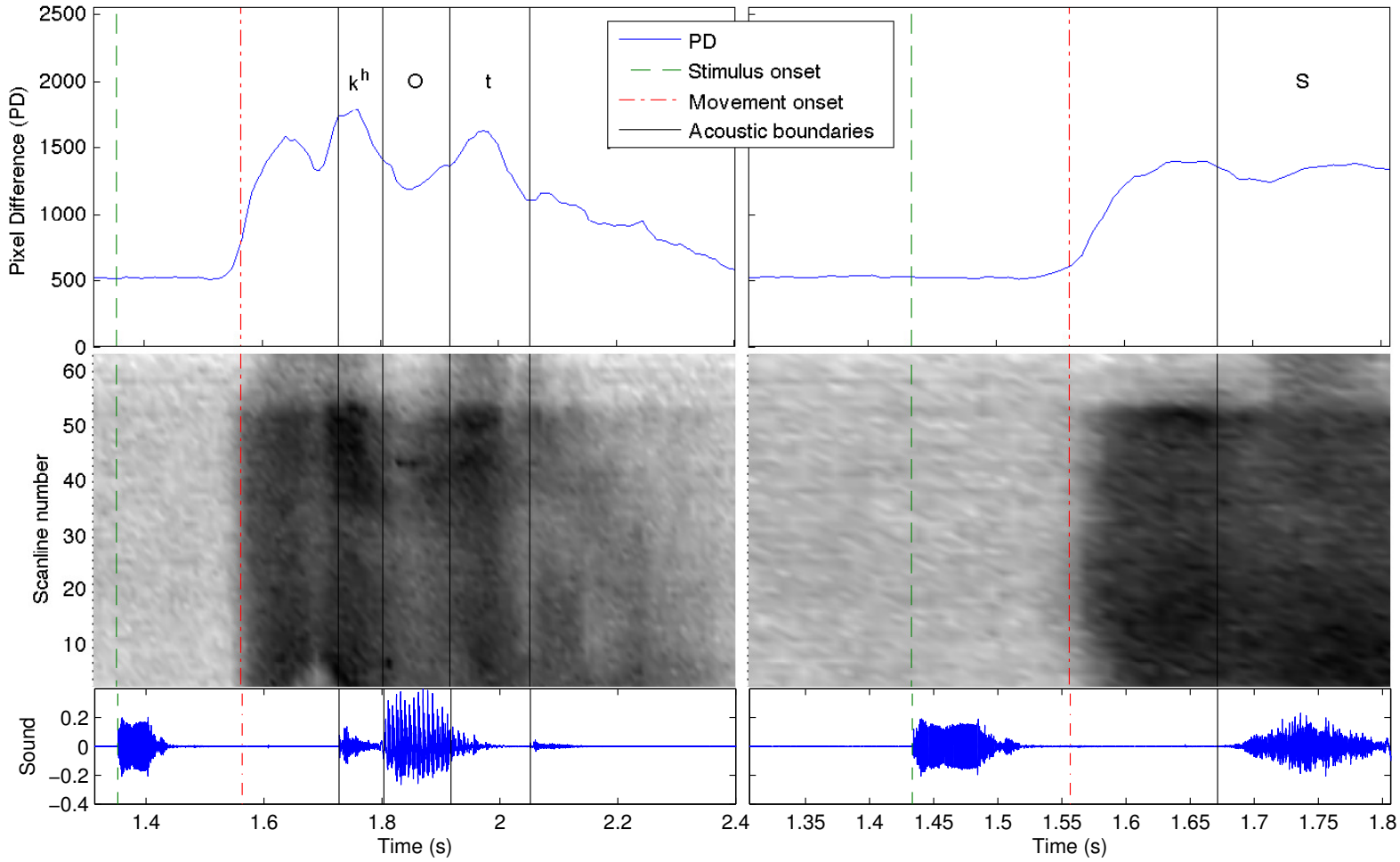

Figure 3.10: Examples of SBPD. The left column has the same recording as Figure 3.4 above. The right column is also from participant P1 in Experiment 2 and shows a zoom of the initiation of [ʃ] in 'sheet'.

3.5 Local articulation onsets

As discussed in Section 2.4.2 ultrasound data contains a lot more actual, utilisable data than usually gets used. The most popular analysis method, splining, ignores everything but the tongue surface. It can be said that PD does something similar. It does after all reduce the image sequence to a sequence of numbers, one per frame pair.

Two ways of expanding PD to utilise the data in a more detailed way are readily evident in the structure of the raw ultrasound data. Instead of asking what the total amount of change is, we could ask what the amount of change is at a given imaging depth or at a given scanline. Of these two options the latter makes more sense for two reasons: First, the raw data is constructed out of scanlines, which means that pixels in a single scanline originated from the same ping-echo-process. Second, it is phonetically more interesting since scanlines correspond to locations along the tongue making them anatomically easier to interpret than scanning depth, which does not even have an easy interpretation when the pixel in question is above the tongue surface and any change in its brightness is produced by indirect echoes within the tongue.

The third option of calculating a pixel-by-pixel change metric is not desirable because it would not reduce the dimensionality of the data. Dimension reduction is essential in this case because it facilitates visualisation of the data and the most salient change in it. It is also desirable because individual pixels contain a lot of spurious change that would interfere with analysis.

A method for identifying local articulation onsets is used for two purposes in this thesis. First, as we will see in Chapters 5 and 6, it gives us further insight into which part of the tongue initiates movement. Second, as shown in the next section, local articulation onsets can be used to implement a fully automated method for detecting the overall articulatory onset.

To identify local onsets in UTI data, a two stage method is used. First, SBPD is calculated for each token. Second, the articulation onsets for each scanline are identified individually with dynamic time warping in the same

manner as described for identifying the articulatory onsets based on PD above in Section 3.3. This makes it possible to identify the tongue regions that initiate movement in a given token.

An example of raw data from the method is shown in Figure 3.11. It shows the distribution of local movement onset latencies for all of the tokens produced by participant P1 in Experiment 2. The figure is divided into individual panels for each of the onset consonants of the target word. Tokens with complex onsets are displayed in a different colour in the panel of the corresponding initial consonant. The data was not thresholded and we can see that a lot of the individual scanline onsets occur before the go-signal.

Already without any aggregation (averaging or medianisation) of the data, it is possible to identify trends in the data, despite the high degree of variation evident in Figure 3.11. In all of the panels the semi-transparent dots representing individual data points cluster in a band whose shape depends on the initial segment. All panels show a band that is bunched to the left in the middle, indicating that scanlines 10-40 register movement onset before those either further front (40-63) or further back (1-10). Most initial segments – with the exception of /d,t,g,k/ and /s/ – show a delay at the anterior scanline numbers ranging roughly from 40 to 60. These scanlines correspond to the area where the mandible shadow is in the images. Of the initial segments that deviate from this general trend, /g,k/ have the flattest distributions indicating what appears to be almost uniform activation across the imaged area from the most anterior to the most posterior scanlines with tongue data in them.

The behaviour of complex onsets does not have a simple relation to how the simple onsets behave. Consider, for example, the complex onset /tr/ (displayed in blue in the panel labelled 't'), which appears to be overall slower than a simple /t/ onset, while the complex onset /bl/ appears to be overall faster than a simple /b/ onset. Meanwhile, many of the panels show considerable overlap of the onset distributions of simple and complex onset tokens and typical behaviour is often difficult to judge from these dense distributions.

To produce a graph of typical local onsets for a given speaker, first the

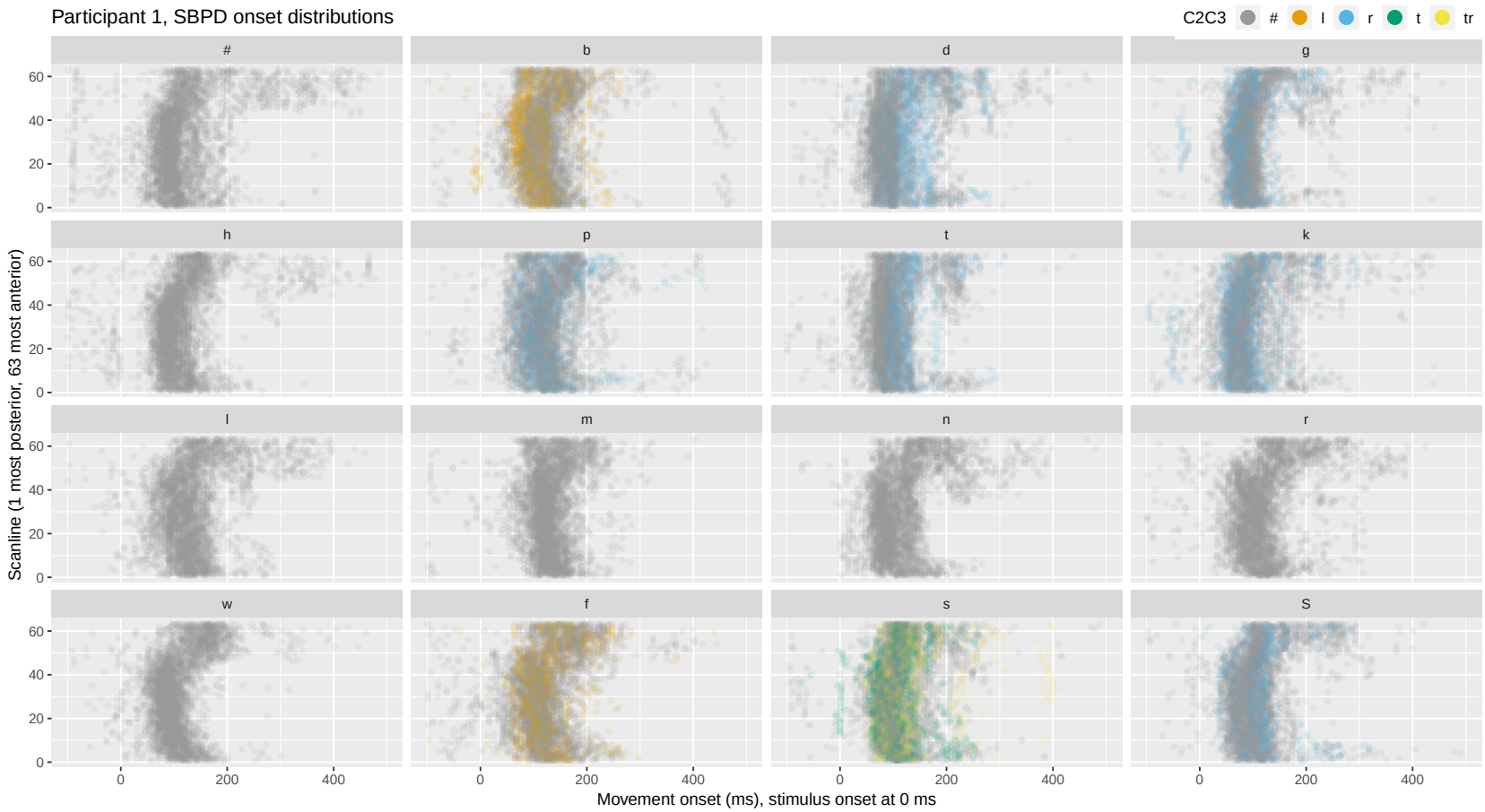

Figure 3.11: Distribution of localised movement onset latencies for participant P1. X-axis is time in milliseconds from stimulus onset and y-axis position from back (1st scanline) to front (63rd scanline) conditioned by initial consonant (# marks no onset, that is, a /VC/ word). Grey dots correspond to /CVC/ utterances and the identity of C1 is indicated on the panel. The coloured dots correspond to utterances with complex onsets and identity C2 and C3 is indicated by the colour.

full distributions of scanline-based onsets were calculated for each participant. Second, the data was thresholded using the 28 ms lower bound on *acoustic* onset time. No thresholding was performed on any other variable including the SBPD data. The justification for the 28 ms lower bound is given above in Section 3.3.2. Finally, the median of the onsets for each individual scanline was calculated within each onset. This procedure gives a view of the average articulatory onset latency as a function of the scanline. The median was chosen as the aggregation function, because it is robust against outliers in the data.

Figure 3.12 presents the result of medianisation of the individual scanline onsets for participant P1 from Experiment 2. We see that in all of the cases recorded and analysed here the first region to move are scanlines 20-40, which sharpens the earlier estimate of scanlines 10-40. A more formal analysis of this is presented in the results of Experiment 2 in Chapter 5. We also see that the delay in the mandible region (scanlines 40-63) is present in all of the panels, but to a varying degree. Especially in labials (second column from left, [b,p,m,f]) the mandible seems to move earlier in this participants case. As we will see, this finding will be important in interpreting the results of Experiment 3 on the differences between UTI and Electromagnetic Articulography (EMA).

Looking at the middle panels in the two top rows, we see that for this participant complex onsets of [b,p] is initiated earlier than a simple onset and that the pattern reverses for [d,t]. This might be a difference in how complex onsets beginning with a bilabial and complex onsets beginning with an alveolar sound are produced by this participant. There is also a more varying degree of onset delay in the scanlines 1-10, but no immediately evident pattern of dependence with the onset phoneme(s).

We should not draw strong conclusions from the data of a single speaker nor even from the larger sample given in Chapter 5. The Figures 5.6 – 5.9 on the whole set of four participants are discussed in that chapter, and in Chapter 6 the ultrasound data of one speaker is related to EMA data from the same speaker. A larger sample of speakers with anatomical measurements is required to properly understand the variation in the localised onset data.

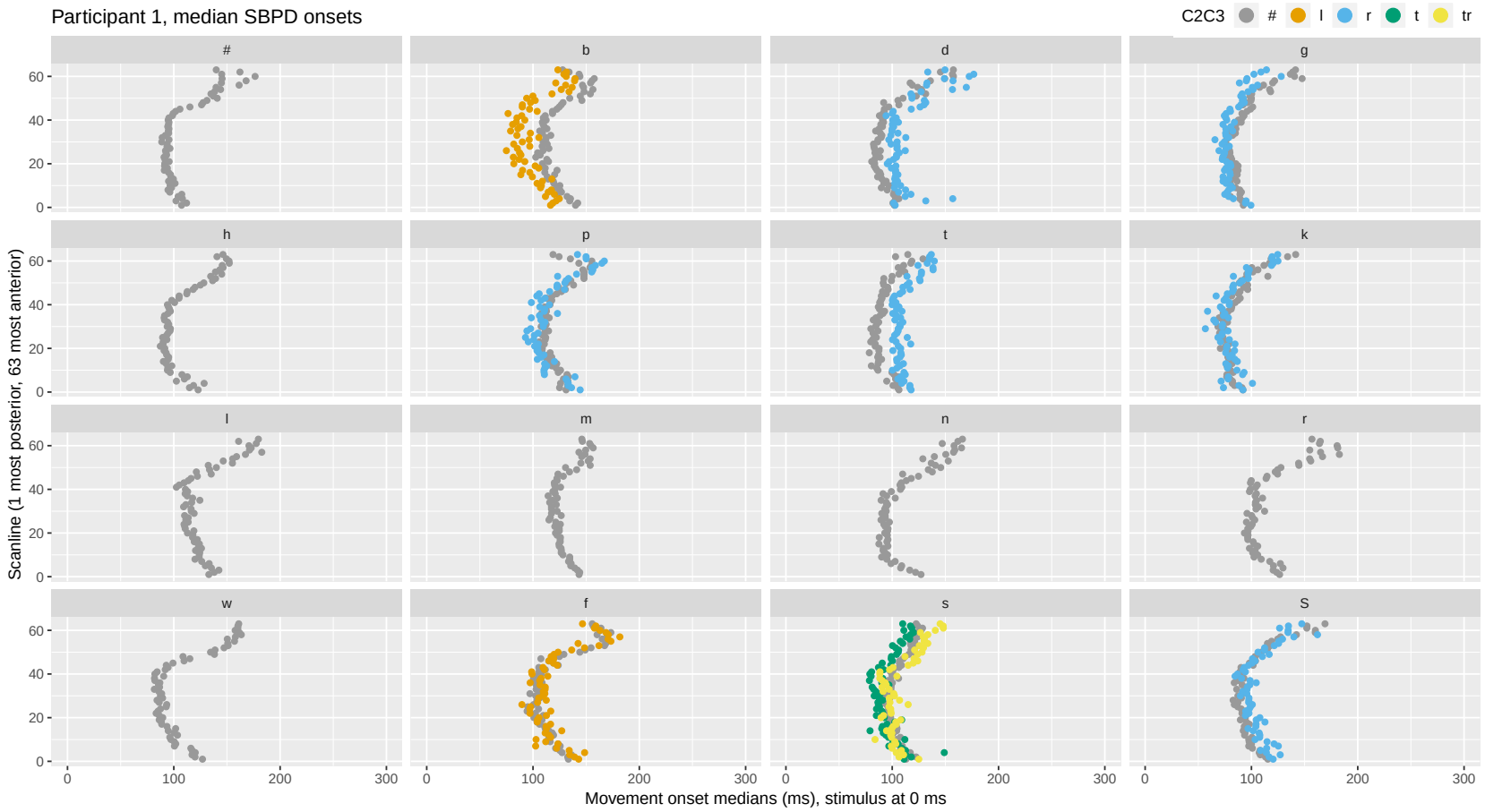

Figure 3.12: Medians of localised movement onset latencies for participant P1 in Experiment 2. X-axis is time in milliseconds and y-axis position from back (1st scanline) to front (63rd scanline) conditioned by initial consonant (# marks no onset, that is, a /VC/ word). Grey dots correspond to /CVC/ utterances and the identity of C1 is indicated on the panel. The coloured dots correspond to utterances with complex onsets and identity C2 and C3 is indicated by the colour.

3.6 Automated onset detection based on SBPD

As is evident above in Section 3.3, the automated onset detection based on PD leaves room for improvement, because its correlation values with both manual video annotation and manual PD annotation, while significant, were not even as good as that of the video annotation with the manual PD annotation. This might be because the dynamic time warping often locks onto a small spurious rise in the signal before the actual articulatory onset or to one of the peaks of the articulatory gestures after the articulatory onset. This section presents an effort to find a better automated onset detection method.

Now, if we look at the detection results in the previous section, we can see that while there is frequent erroneous early triggering present in individual scanline-based onsets, the majority of the local onsets have very reasonable values. Furthermore, with aggregation we can produce robust results such as those in Figure 3.12.

There is a way of using the principle of aggregation in the analysis of a single token, by taking the median over the scanline-based onsets of the token. In effect, we are letting the scanline-based onsets vote: instead of selecting the first – most likely too early – triggerings to be our measure of movement onset, we are selecting the point at which half of the scanlines have triggered and half are yet to trigger. This produces a more robust way of automatically determining movement onset.

The SBPD movement onsets were calculated in Matlab with dynamic time warping on individual scanlines as described in the previous section. After this the data was loaded to R and aggregation was performed by taking the median over the scanlines for each recording.

3.6.1 Data loss

The same test set of 333 tokens from Experiment 2 was used as earlier in Section 3.3. Only two tokens had an automatic SBPD based onset time that was below 28 ms. So, data loss due to use of this method is negligible in this test

set. Excluding these two tokens and running the same thresholding and exclusion operations as earlier in Section 3.3.2, we are left with 231 tokens with full observations.

3.6.2 Correlations of the SBPD annotation method with others

The results of the second annotation experiment are presented in Figure 3.13, which shows scatter plots of the SBPD based automatic annotation and of the PD based automatic annotation against the manual video annotation and the human PD annotators. We can see that overall the SBPD and PD based automated methods produce similar results. However, there is more scattering apparent in the automated PD results (panel b) when compared with the automated SBPD results. Indicating that the SBPD method produces fewer outliers and can therefore be considered more robust. This is confirmed by the correlation values in Table 3.3. All correlations test as statistically significant with P-values $< 2.2 \times 10^{-16}$ when tested individually against the null hypothesis that $\rho = 0$ using a two-tailed t-test with 229 degrees of freedom.

The level differences among the annotation methods were studied like in Section 3.3.3 with ANOVA analysis followed by Tukey's honest significance test. The ANOVA F-statistic was 49.678 on 6 and 1610 degrees of freedom resulting in

Table 3.3: Pearson's correlation coefficients of manual ultrasound video labelling, automated articulatory onset detection based on SBPD and PD, and manual labelling of the PD curves by three human judges. Judge one (the author) labelled the data set two times (days 1 and 2).

	Manual	Automatic		Manual PD annotation			
	Video	SBPD	PD	Day 1	Day 2	Judge 2	Judge 3
Manual Video	1	0.826	0.657	0.817	0.789	0.750	0.736
Auto SBPD	0.826	1	0.821	0.811	0.809	0.791	0.779
Auto PD	0.657	0.821	1	0.690	0.675	0.660	0.650
Judge 1, Day 1	0.817	0.811	0.690	1	0.952	0.914	0.912
Judge 1, Day 2	0.789	0.809	0.675	0.952	1	0.920	0.920
Judge 2	0.750	0.791	0.660	0.914	0.920	1	0.895
Judge 3	0.736	0.779	0.650	0.912	0.920	0.895	1

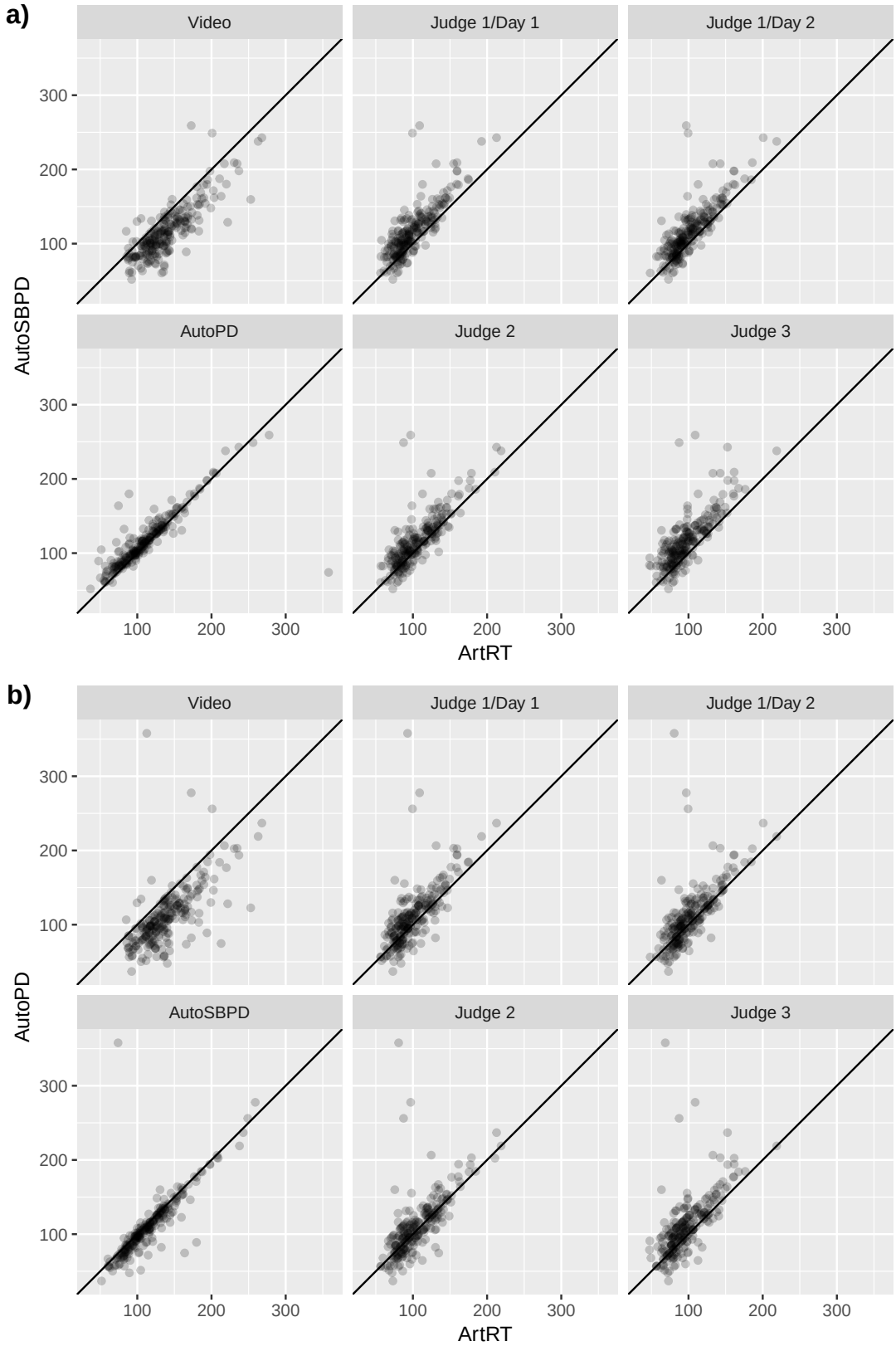

Figure 3.13: Automatic SBPD and PD onsets compared with manual video annotation and manual PD annotation. The panels show also an identity line ($y = x$) to make it easier to see the general trends in the data. Scales are in milliseconds.

a P-value < 2.2×10^{-16}. The results of the Tukey test are presented in Figure 3.14. The new comparison results on the automatic SBPD method are displayed on lines 1 and 7-11. The old results on the correspondences of the other methods are provided for reference. We see that in relation to all of the other methods automatic SBPD onsets are closer to the manual video annotation results. At the same time they are statistically not significantly different from the automatic PD onsets, and is further removed from the manual PD annotation results than the automatic PD onsets. Combining these results with the generally higher correlations means that automatic SBPD onsets should be preferred over automatic PD onsets.

3.7 Summary

Pixel difference (PD) shows promise as a method of visualising change in ultrasound videos. The analysis on which time step to use, concluded that for purposes of analysing speech articulation a time step of 1 is ideal when using sampling frequencies comparable to the 120 frames per second (fps) used in the data of this thesis. If using higher sampling frequencies, this issue should potentially be revisited. If using lower sampling frequencies, the user should be aware of the possibility of the pulse affecting the contours in a more significant way than it does with the current frame rate.

Scanline-based pixel difference (SBPD) also shows potential for visualisation purposes, but it requires more effort in interpretation than PD because SBPD provides a more detailed, and therefore, more complex representation of the video being analysed. When used, it should be used in conjunction with the corresponding PD display.

To our knowledge, the localised onsets produced from SBPD are the first evaluation of speech onset location on the tongue. The localised onset distributions in Figure 3.11 and the medianised data in Figure 3.12 provide a novel way of visualising the onset as a process that happens over time. More careful statistical analysis of these patterns can be found in Experiments 2 and 3 in

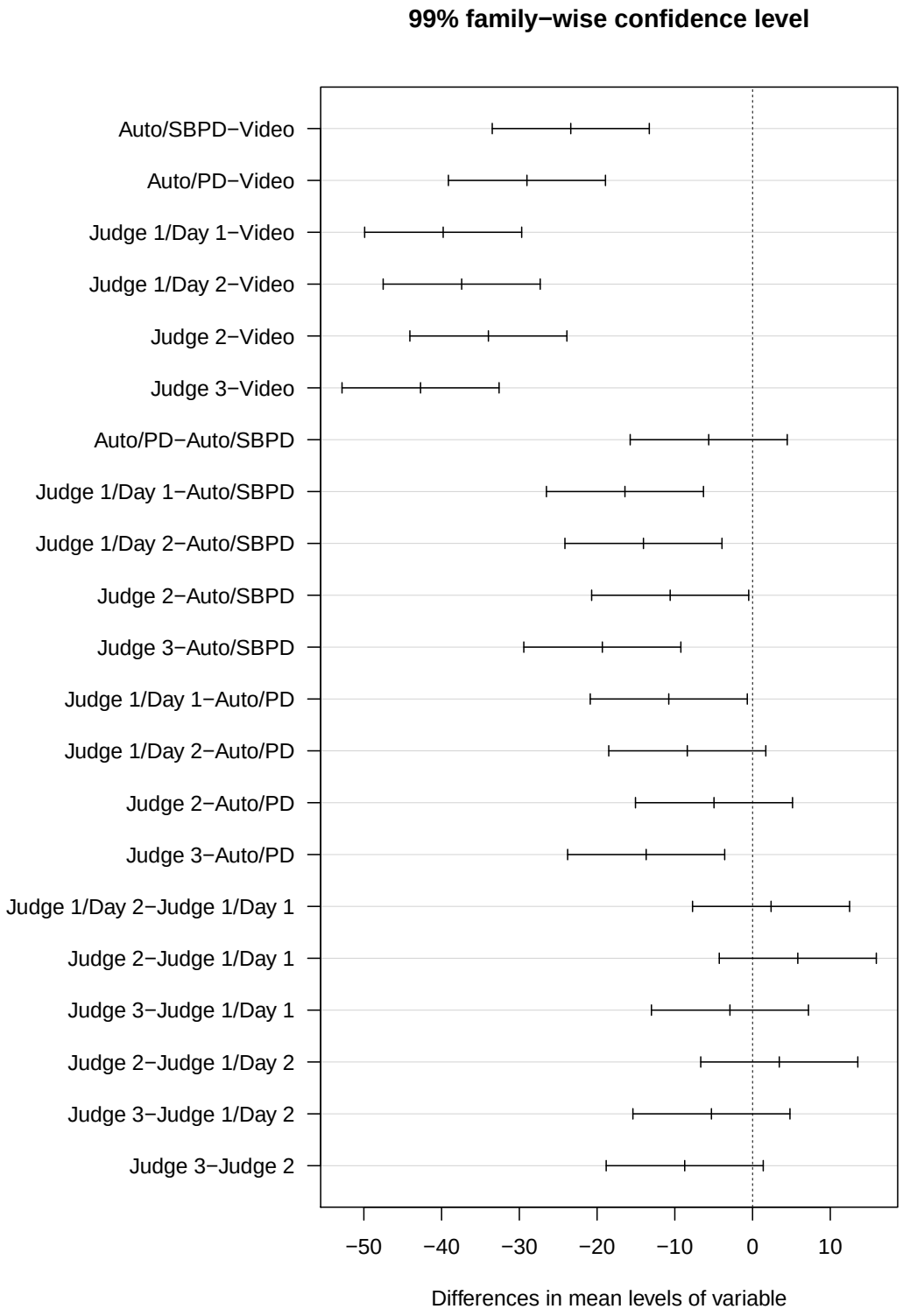

Figure 3.14: Confidence intervals ($P = 0.99$) for post-hoc pairwise comparisons of SBPD based onset detection against the methods discussed in Section 3.3. Methods can be considered to have the same expected value if the confidence interval of their mean difference overlaps 0. Scale is in milliseconds.

Chapters 5 and 6.

Out of the two automated articulatory onset detection methods, onset detection on SBPD is the more reliable since it has higher inter-method correlations than automated onset detection on PD. The general choice of onset detection method should take into account also the resources available. Manual onset detection on PD produces fewer outliers than the automated SBPD method, but requires significantly more work hours than the SBPD method. Considering the rather large data sets analysed, in this thesis onset detection will be mainly performed with the automated SBPD method to have an optimal balance of reliability and efficiency.

Chapter 4

Experiment 1: Picture naming in UTI

4.1 Introduction

This experiment uses an articulatory variant of the Snodgrass-Vanderwart picture naming task with coloured pictures (Snodgrass and Vanderwart 1980, Rossion and Pourtois 2004). The participants named pictures shown on the screen as fast as possible while their speech sounds and articulation were recorded. Speeded picture naming is a useful paradigm for eliciting hesitations and other pre-speech behaviour. The articulation was recorded with Ultrasound Tongue Imaging (UTI). An in-depth technical background for this method and the analysis methods developed for it can be found in the previous chapter.

The data has been gathered for a project predating this thesis work (Schaeffler et al. 2014; 2015), where it was part of a study of hesitations and other non-linguistic movements that occur before acoustic speech onset. For the purpose of the current project, the data from this experiment has been used as a test bed for early development of analysis methods of UTI data before data that was specifically designed to answer the main research questions was available for analysis. The analysis of this data set presented here is an exploration of the way Pixel Difference (PD) represents change as a function of time and is mainly qualitative.

4.2 Materials and methods

4.2.1 Participants

Five participants were recorded: four females (E1, E2, E3, and G1) and one male (F1). All of the speakers used their first language in the experiment: three in English (participants E1, E2, and E3), one in German (participant G1) and one in Finnish (participant F1).

Experiment 1 already had ethical approval as it had been acquired by the experiment's principal investigator Dr. Sonja Schaeffler according to the guidelines in place at Queen Margaret University. The data in this project was acquired from volunteer participants. For general ethical principles the participants had the opportunity to stop the experiments at any point. This is especially important when using UTI with probe stabilisation such as the headset used in these experiments, because wearing the headset can be strenuous in longer experiments. The data was anonymised at time of recording.

4.2.2 Procedure

The experiment was run at the UTI facility at Queen Margaret University. The display of stimuli and recording of ultrasound data, lip videos, and audio was controlled with Articulate Assistant Advanced (AAA) software (Articulate Instruments Ltd 2012). The probe was stabilised with the headset shown in Figure 2.14 (Articulate Instruments Ltd 2008). The participant sat in front of a computer screen, that was used to display the visual stimuli of the experiment, in a recording booth fitted with sound proofing wall materials and the ultrasound equipment. The experimenter sat in a control room where they could trigger the experiments and monitor them with sound as well as ultrasound and video displays.

The participants were asked to name out loud picture stimuli which were shown to them on the computer screen in front of them. The pictures were coloured versions of the pictures of the Snodgrass picture set (Rossion and

Pourtois 2004, Snodgrass and Vanderwart 1980). Three examples of the original uncoloured pictures are shown in Figure 4.1. The pictures were displayed in random order. Before the recordings, the participants had been given the instruction to name the pictures as soon and as accurately as they could.

Each trial was began when recording was initiated by the experimenter. This started the sound recording and caused a fixation point to be shown on the participant's computer screen for 1.5 seconds before the picture stimulus appeared on the screen. The synchronised ultrasound recording was automatically initiated at about 0.5 s or 1.0 s after the sound recording began thus capturing any movements related to speech preparation as well as making it possible to spot cases where the subject was moving already before the onset of the stimulus. The delay results from the experimental software and the experimental apparatus and can not at the present be controlled, but can be and has been carefully measured for each token by the recording system.

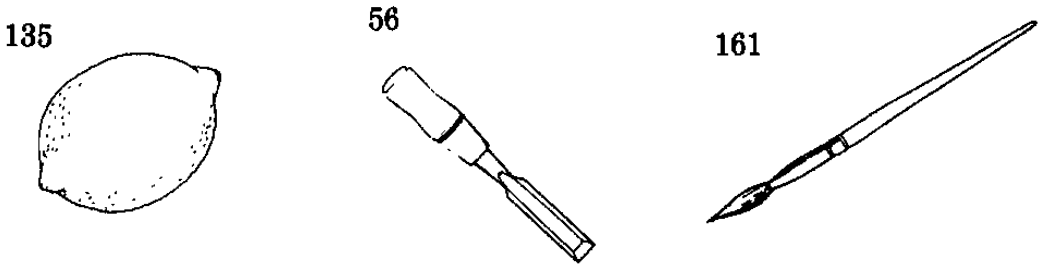

Figure 4.1: Three examples of the original Snodgrass pictures: a lemon, a chisel, and a paintbrush. The experiment used coloured versions of the pictures.

4.2.3 Audio and UTI recordings

The experiment was run with synchronised ultrasound, lip imaging video, and sound recording controlled with the AAA software (Articulate Instruments Ltd 2012). The participants were fitted with a purpose-built headset to ensure stabilisation of the ultrasound probe (Articulate Instruments Ltd 2008). Attached

to the helmet was a small Audio Technica AT803b microphone for high-quality acoustic recordings. Ultrasound recordings were obtained with the high speed SonixRP system at Queen Margaret University (Wrench and Scobbie 2011). Ultrasound frame rate was 201 frames per second (fps). Each frame consisted of 38 scanlines over a field of view of 115.4°.

4.3 Audio analysis

Unlike in the two following experiments, the audio data for this experiment was not phonetically segmented. The data in this experiment is not phonetically balanced, and we do not have a reasonable guarantee that the participants actually produce the intended target words for each picture. Furthermore, picture recognition and lexical retrieval and other planning processes will be part of both articulatory and acoustic reaction times in this experiment confounding any effect that might be due to phonemic content of the recorded utterances. As a result, the purpose of audio analysis in this experiment was only to provide acoustic onset times of the target words for testing different time alignments of PD (see Section 4.5).

Acoustic onsets of the target words were labelled by the author in Praat (Boersma and Weenink 2010) based on the waveform and the spectrogram with the default settings, with the exception that the spectogram's frequency range which was set to go up to 10 kHz. Since the data includes a great number of non-speech sounds as well as vocalisations of uncertainty (such as 'Ummm, I don't know'), it was not always clear what should be defined as the actual acoustic onset. In such cases listening to the audio recording was used to help with the judgments, but even then, some of the tokens were very difficult to analyse.

4.4 Articulatory analysis

First, PD contours were computed for all of the tokens. Second, the author manually labelled the articulatory onset on the PD contours of each token. The tools for these analysis steps are described in Chapter 3 and the Matlab code for them is included in Appendix B.

To map out variation in the data a qualitative analysis of the data was performed by examining the PD graphs of the whole data set. The author also performed a qualitative analysis of the tokens by first examining their PD contours until he was familiar with the whole data set. Three categories were identified in the tokens where the participant had produced a speech response: 'steady', 'hesitation', and 'chaos'. Tokens where the participant did not speak, were assigned the category 'no speech'. The categories are described in more detail and examples are given in the next section.

In the second phase of the qualitative analysis, the author re-examined each token's PD contour and assigned them to one of these categories. The results of this analysis are reported in Section 4.5.

4.4.1 Stability categories of picture naming tokens

In the recordings where speech was present, three distinct categories of production plus one category without speech were found. The production categories are illustrated by examples in Figures 4.2 – 4.4 along with the corresponding sound waveforms. First, Figure 4.2 shows a typical steady production. The participant has held still before starting speech articulation.

Second, Figure 4.3 shows an example of a hesitation. The participant moves her tongue as if she was going to speak, but returns to rest and finally speaks later. There is often sound associated with these types of hesitation. The hesitation sounds range from brief sounds of the mouth opening (for example, clicks and smacks) to long vocalisations with or without meaningful content (for example, 'Errrr', 'Ummm', 'Ummm, I don't know. . . '). The defining feature of this type of pre-speech behaviour is that the participant returns to rest before

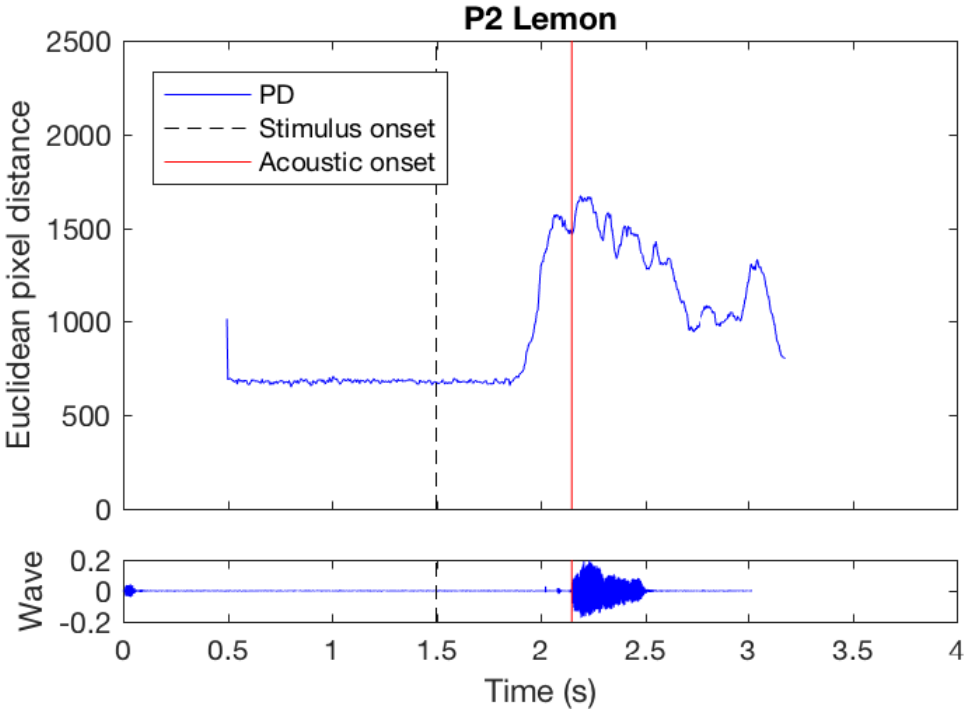

Figure 4.2: A steady production: E2 naming a picture of a lemon.

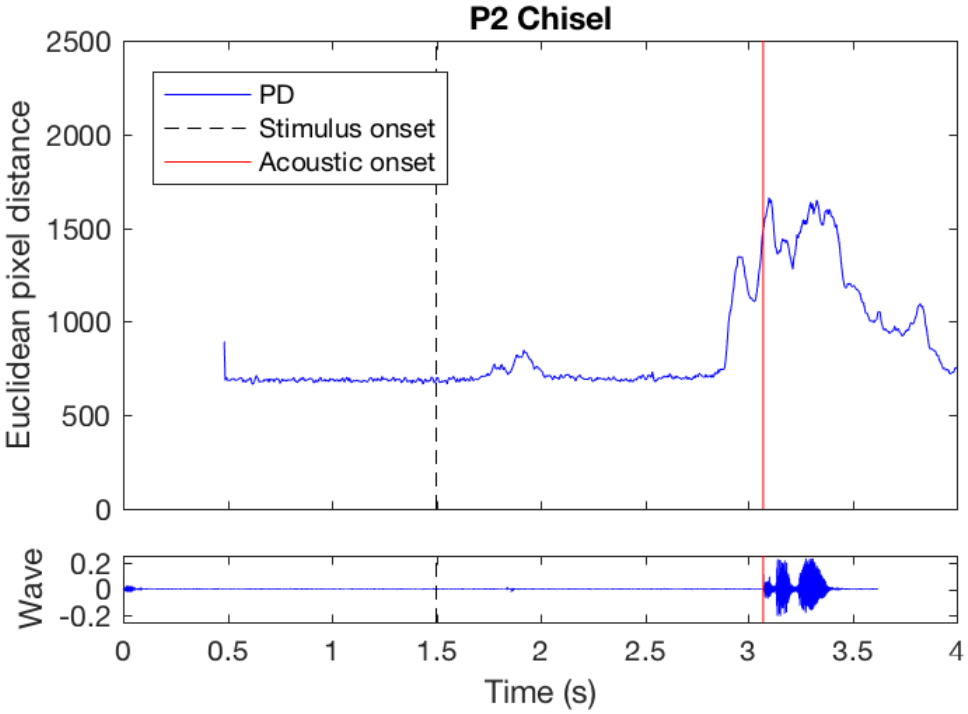

Figure 4.3: A hesitation: E2 naming a picture of a chisel.

pronouncing the target word.

Third, Figure 4.4 shows a chaotic example. The participant is moving already at the time the recording starts and continues to move throughout the whole recording. This is a very varied class with some tokens having an ongoing vocalisation accompanying the movement, while in some the participant moves silently until they produce the target word making it difficult to correctly categorise these tokens based on only acoustics.

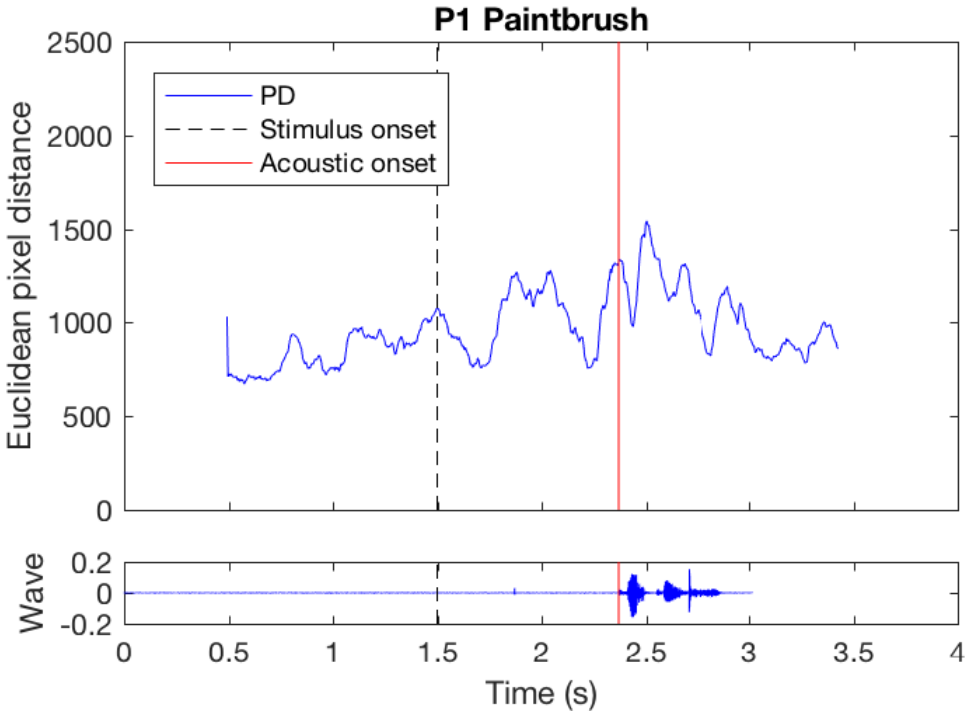

Figure 4.4: Chaos: E1 naming a picture of a paintbrush.

The fourth category – 'no speech' – is not illustrated because these were recordings where the participant failed to respond within the recording time. Since there was no speech to analyse, they were excluded from further analysis.

In most tokens the pre-speech movements are relatively small in comparison to the speech movement as seen in the examples of 'steady' and 'hesitation' Figures 4.2 and 4.3. In contrast, as we see in 'chaos' in Figure 4.4, in this category the movements are of the same magnitude at all stages of the production.

4.5 Results

The results from PD analysis reveal variation and systematicity in the productions. The greatest differences occur between speakers, but individual speakers also display great variation. Mainly qualitative analysis is used in exploring this variation below, but it is augmented with some quantitative analysis as well.

4.5.1 Descriptive statistics

Only two of the participants completed the whole experiment of naming 260 tokens. Wearing the UTI helmet gets strenuous in longer experiments and naming all 260 tokens takes about 40 minutes. These factors resulted in varying number of tokens being recorded for each participant. The number of recorded tokens for speakers E1, E2, E3, F1, and G1 (before removal of missed trials) was 148, 264 (includes four repeated trials), 219, 216, and 260 (mean = 221) recorded tokens, respectively. However, all of the recorded tokens did not contain analysable speech. The sample sizes given in Tables 4.1 and 4.2 are the numbers of tokens that were analysable in the table's modality.

Descriptive statistics of the data set are listed in Tables 4.1 and 4.2 for acoustic and articulatory reaction time, respectively. Tokens that had an onset time before stimulus onset (negative reaction time) have been removed separately for each modality resulting in slight differences in the sample sizes used.

Notable features include differences in average reaction time – F1 is clearly over all slower to respond than E1 or E3; he also has the greatest range of acoustic

Table 4.1: Minimum, mean, standard deviation and maximum (all in ms) of the acoustic reaction time of each speaker.

Speaker	min	mean	sd	max	n
E1	424	662	231	1757	152
E2	300	970	392	3865	260
E3	494	807	317	2846	216
F1	1092	1424	377	5010	216
G1	410	1018	505	4028	256

Table 4.2: Minimum, mean, standard deviation and maximum (all in ms) of the PD based articulatory reaction time of each speaker.

Speaker	min	mean	sd	max	n
E1	4	324	168	1296	146
E2	3	560	444	2099	232
E3	108	421	278	2286	217
F1	786	985	142	2227	215
G1	212	683	398	3103	253

onset times. G1 stands out as almost as slow as F1 in acoustic response times but produces faster articulatory response times and has the greatest range in them of all of the participants. It is also of note is that in terms of standard deviation the participants rank in acoustics as E1 << E2,E3,F1 << G1, but in articulation as F1,E1 << E3 << G1,E2.

4.5.2 Variation in pre-response stability

Variation in pre-response stability was investigated with a qualitative categorisation analysis of PD contours. The analysis procedure and the categories found in the data are described in Section 4.4. Frequencies of the different token types for each speaker are listed in Table 4.3 and the relative proportions are illustrated in Figure 4.5. Apart from speaker E2, all of the speakers produce a high proportion of steady productions, some hesitations and few chaotic recordings. E2 stand out from the rest by being the only speaker with less than half of her tokens in the steady category.

To test this apparent dependence of proportions on speaker identity statistically, we ran a χ^2 test of proportions. The 'no speech' category was excluded from the analysis because of too low expected number of tokens in cells corresponding to it. The test shows that proportions of types depend on the speaker in a statistically significant manner: χ^2 = 203.75, degrees of freedom = 8, and P-value < 2.2×10^{-16}.

Table 4.3: Frequencies of the production categories in the data of each speaker. Examples of Steady, Hesitation, and Chaos are shown in Figures 4.2-4.4. 'No speech' refers to trials where the participant failed to respond and did not speak.

Type	E1	E2	E3	F1	G1
Steady	94	91	170	174	201
Hesitation	38	85	34	39	46
Chaos	22	84	16	3	10
No speech	1	4	1	0	3
Total	155	264	221	216	260

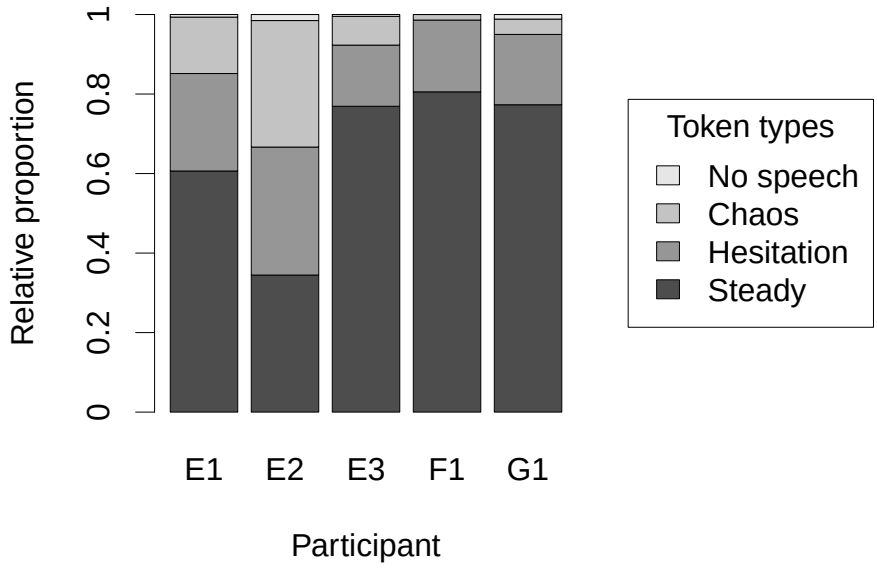

Figure 4.5: Relative proportions of the different token types in Experiment 1 for each participant.

4.5.3 Pixel difference based onset distributions

Figures 4.6 – 4.10 are PD plots of all of the tokens with a detectable articulatory onset for each speaker. The onsets are from the manual labelling of the PD contours as described above in Section 4.4. The plots are divided into three panels with the top panel displaying the PD contours time aligned at three different points. In each panel the time axis is adjusted so that the alignment point is at $t = 0$ s.

The top panel is aligned at stimulus onset – the moment when the picture to be named appeared on the computer screen. The middle panel is aligned at the articulatory onset time determined manually from the PD curves. Finally, the bottom panel is aligned at the acoustic onset time.

The base level of PD, that is, the noise floor of PD, is visible on the left side of all of the panels as the dark opaque band at about 750 units. This is followed by the steep rise in the curves as speech articulation begins. On the right-hand side we see that, as speech articulation ends, most of the curves settled back down to the base level that preceded articulation, but with considerably more random variation evident in the haze formed by some of the curves remaining at higher levels of activation.

As one would expect, the curves are out of synch in the top panel. This is a result of variation in articulatory reaction time between the trials. The middle panel is the most focused, because in it the contours have specifically been synchronised on the rising edge of articulatory onset. In the bottom panel, the distributions move out of focus again indicating that acoustic onsets do not fall on a specific feature in the PD contours. There is, however a constant feature in the bottom panel of each speaker: at the acoustic onset, that is, at $t = 0$, and immediately after it very few of the curves have low-lying trajectories. This is an indication that speech articulation constantly produces high PD values.

The individual characteristics of each speaker that were already evident in the descriptive statistics are corroborated in these figures. E2 has the greatest amount of chaos in her productions. F1 and G1 show a tendency to have longer acoustic response times than the other speakers.

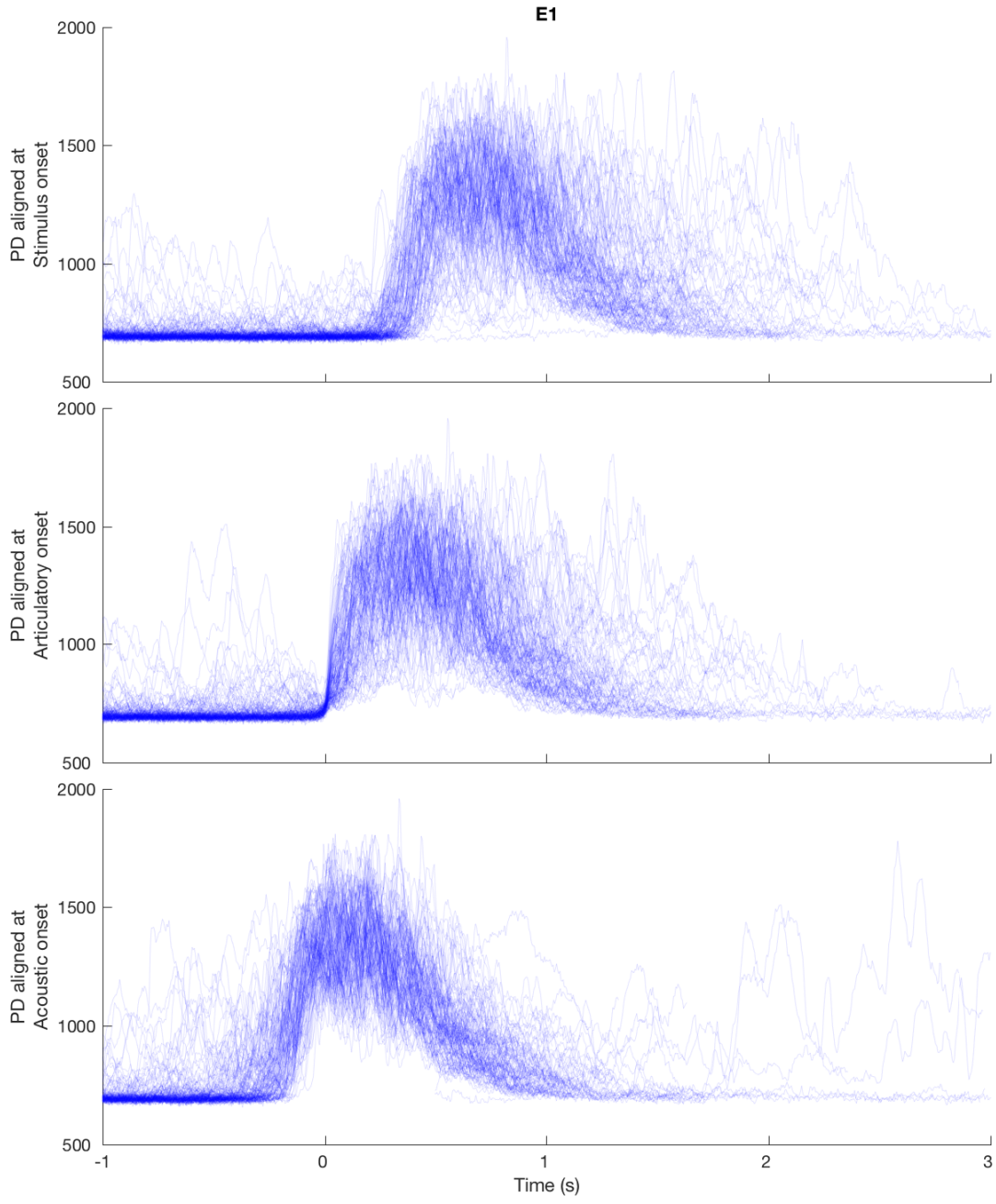

Figure 4.6: PD curve distributions for speaker E1. The curves are time aligned at a different time point in each panel. In the top panel stimulus onset (picture appears on screen) is $t = 0$ s. In the middle panel $t = 0$ s is articulatory onset time determined from PD. And in the bottom panel $t = 0$ s is acoustic onset time.

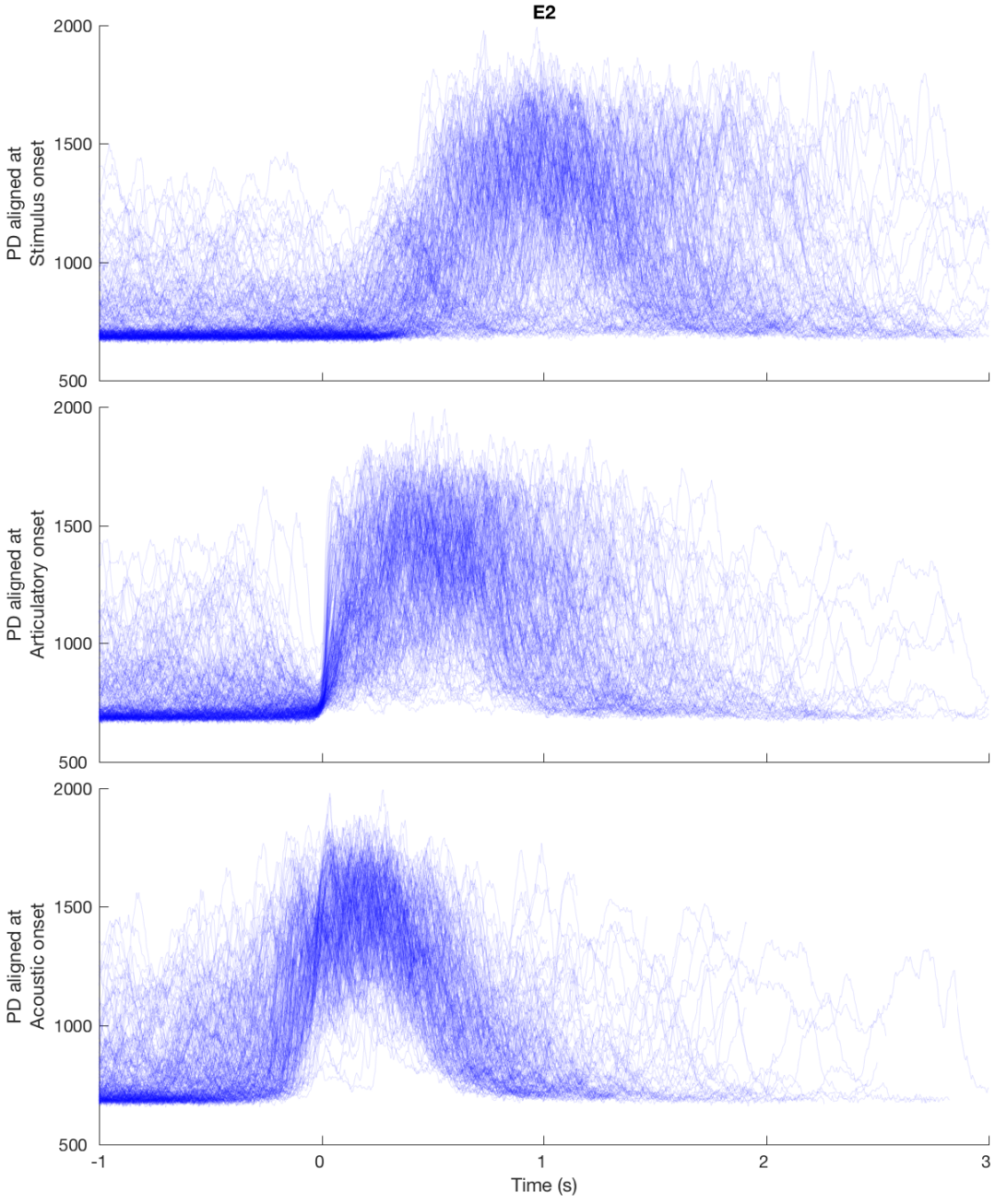

Figure 4.7: PD curve distributions for speaker E2. The curves are time aligned at a different time point in each panel. In the top panel stimulus onset (picture appears on screen) is $t = 0$ s. In the middle panel $t = 0$ s is articulatory onset time determined from PD. And in the bottom panel $t = 0$ s is acoustic onset time.

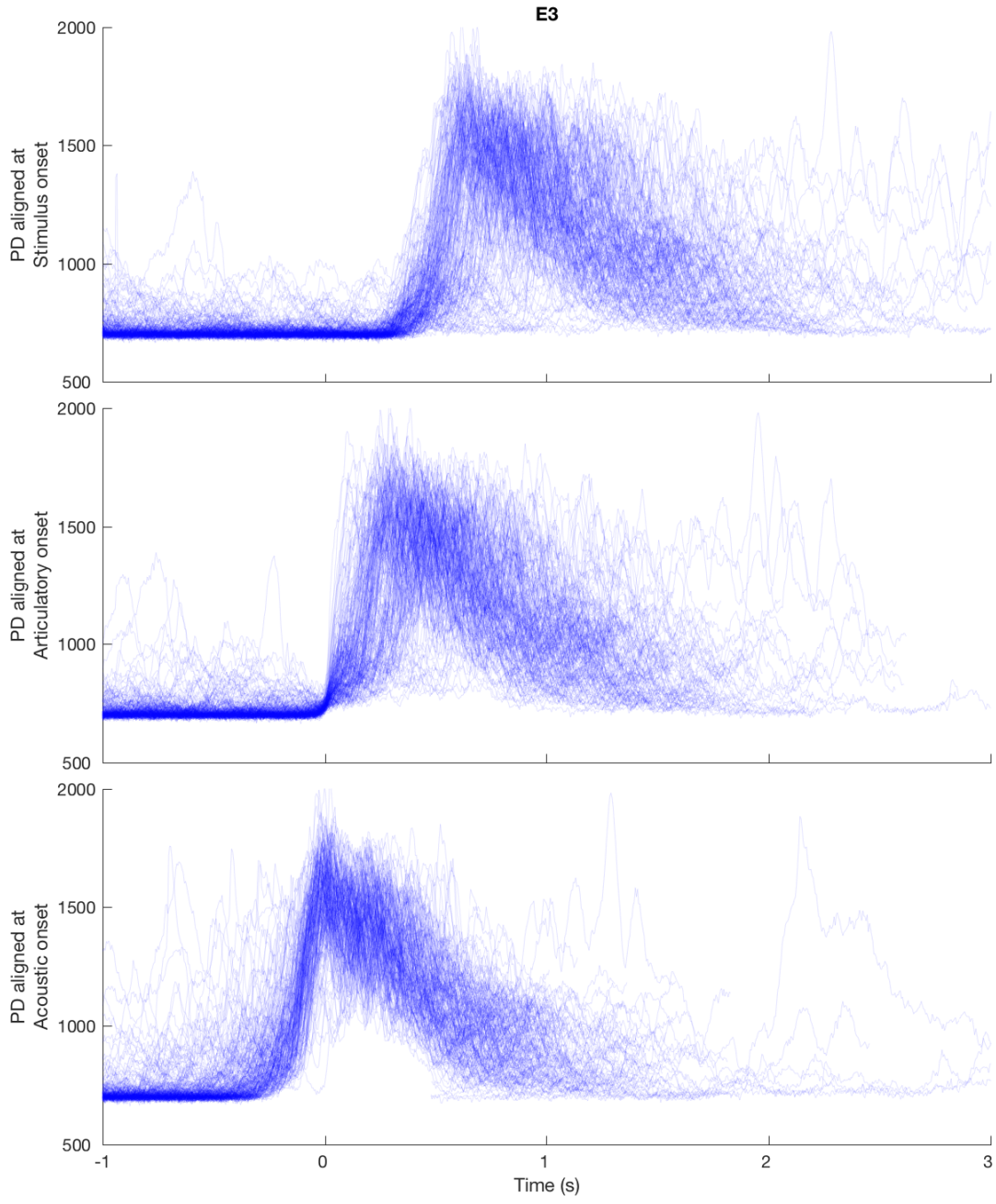

Figure 4.8: PD distributions for speaker E3, with the contours aligned at stimulus onset (top), articulatory onset (middle), and acoustic onset (bottom). The alignment point is at t=0 in each panel.

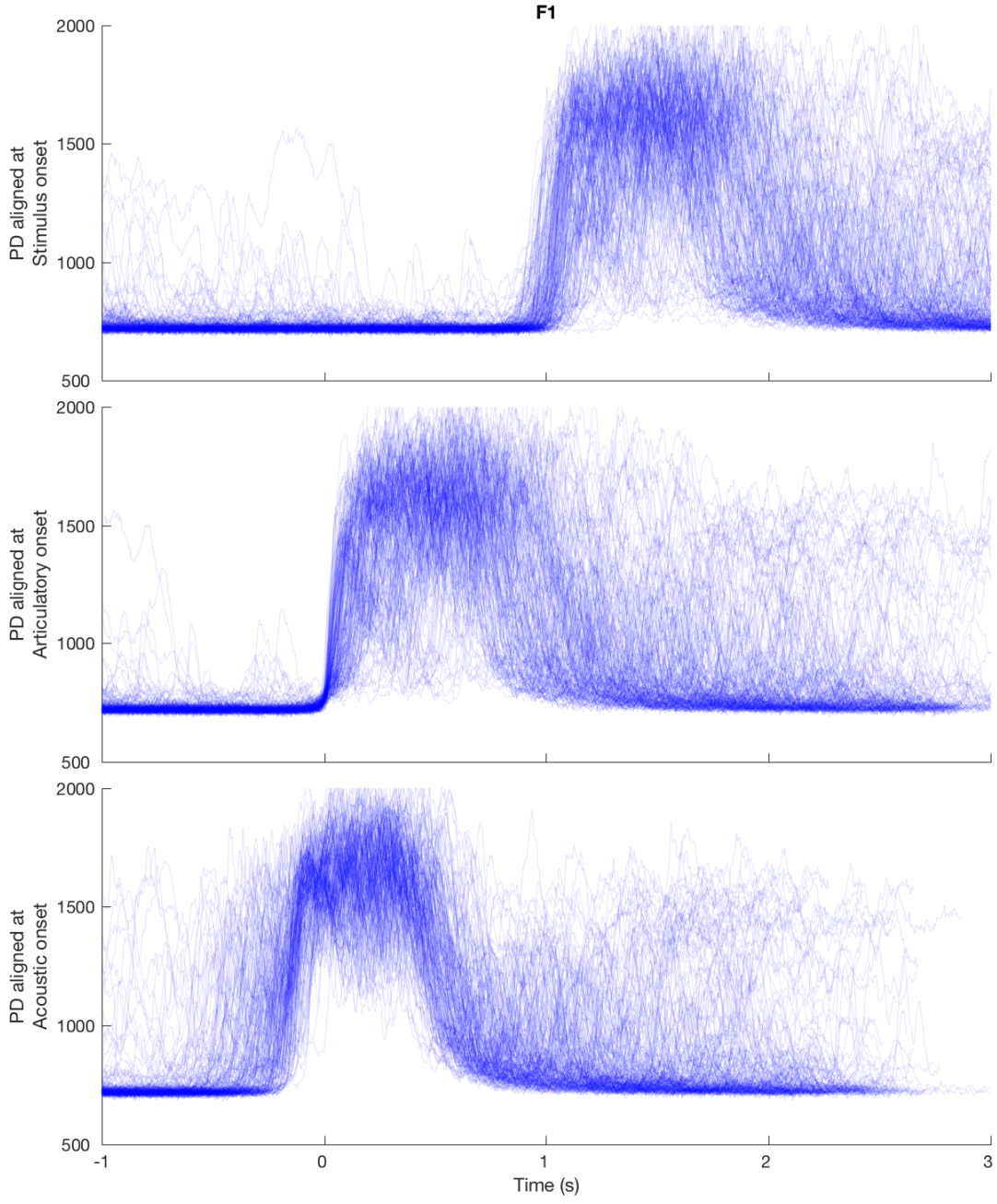

Figure 4.9: PD curve distributions for speaker F1. The curves are time aligned at a different time point in each panel. In the top panel stimulus onset (picture appears on screen) is $t = 0$ s. In the middle panel $t = 0$ s is articulatory onset time determined from PD. And in the bottom panel $t = 0$ s is acoustic onset time.

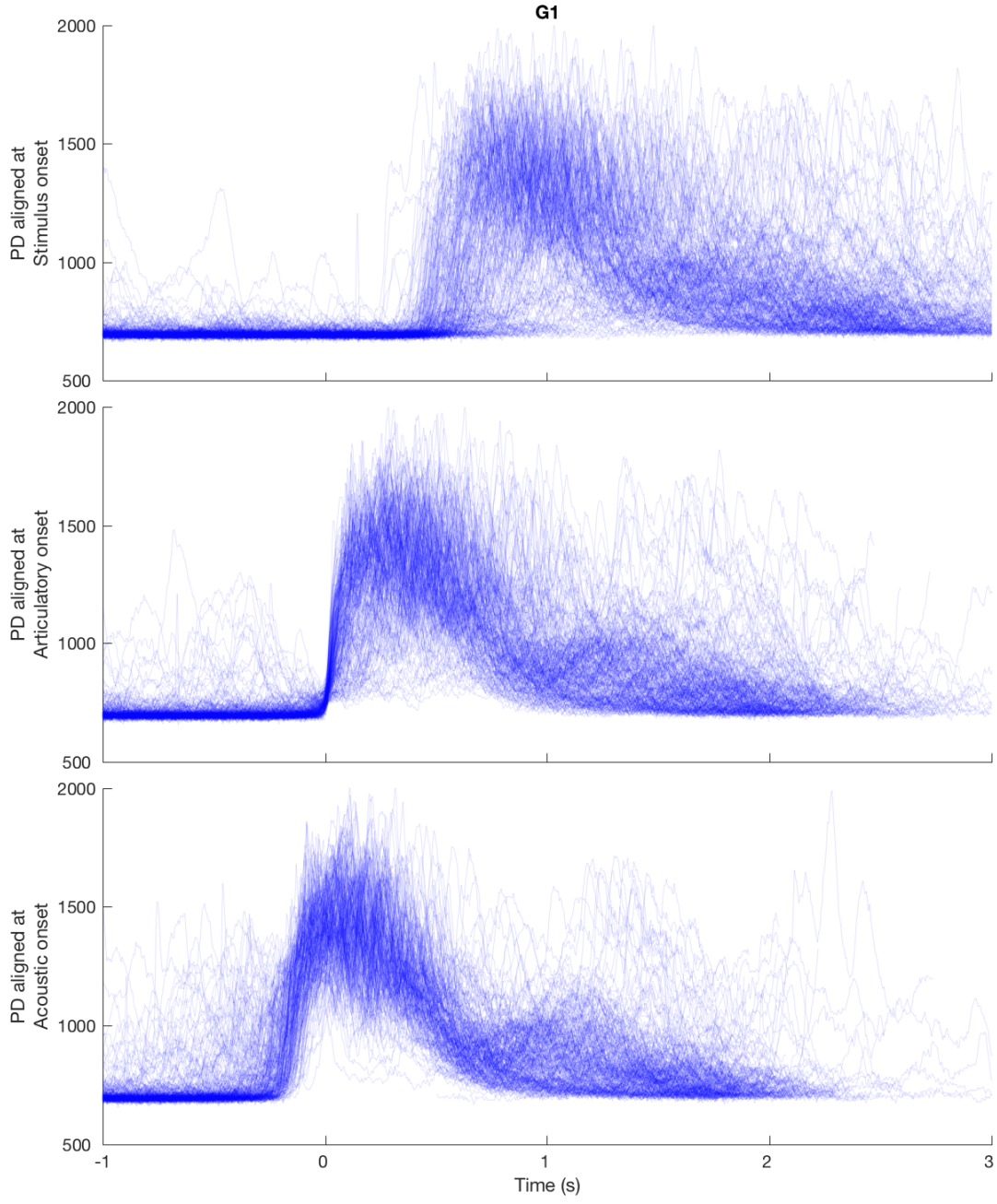

Figure 4.10: PD curve distributions for speaker G1. The curves are time aligned at a different time point in each panel. In the top panel stimulus onset (picture appears on screen) is $t = 0$ s. In the middle panel $t = 0$ s is articulatory onset time determined from PD. And in the bottom panel $t = 0$ s is acoustic onset time.

The middle panels offer evidence of individual differences in how fast articulation begins. Most speakers show a strong rise visible as the baseline band turning up and then dispersing. However, in the case of E3 the baseline band does not turn as steeply up as it breaks up. This means that she has a number of tokens, which do not start as steeply as most of the other speakers' tokens do at the articulatory onset.

As evidenced by the statistics in Tables 4.1 and 4.2, there is considerable inter-speaker variance in both acoustic and articulatory onset times. Looking at Figures 4.6 – 4.10, the only panel which repeats this extent of inter-speaker variance is the top one. This means that the main source of the inter-speaker variation is the articulatory onset latency with acoustic onset time following relatively soon after the articulatory onset time.

4.6 Discussion

A recurring problem in analysing the data of this experiment was how to demarcate between speech and non-speech. The problem exists in parallel in both articulatory and acoustic modalities and also spans them. It is a threefold problem of when does movement cross over from other types of movement to speech articulation, as well as when does vocalisation become speech, and if this boundary between non-speech and speech should be defined primarily based on one of the modalities or both.

In answer, there is a continuum from completely steady productions where the onset of articulation is absolutely clear to productions where it is next to impossible to tell – at least with the current analysis tools – when seemingly random movement turns into purposeful speech articulation. On the acoustic side the problem is often more clear cut, but there were still tokens with vocalisation followed by non-intelligible speech, followed by the target word.

Picture naming – the task in this experiment – is well suited for generating hesitations, false starts, respiratory noises – especially inhalation sounds, and other pre-speech phenomena found prior to the production of the target

word. Frequent hesitations in the data make it useful for its original purpose of studying variation in preparatory behaviours (Schaeffler et al. 2014).

However, manually identifying hesitations in ultrasound videos is challenging. While a more precise analysis of the nature of the hesitations and other extraneous movements is outwith the scope of this thesis, the categorisation experiment shows that using the PD curves as a visualisation tool makes it possible to quickly identify tokens that contain hesitations or other extraneous movements.

Proportionally, for most speakers, the number of tokens in the 'steady' category is greater than those in the 'hesitation' and 'chaos' categories, so being able to easily decide which tokens should be more carefully analysed can be considered a major gain in the efficiency of analysis. Furthermore, having a visualisation method for identifying hesitations at a glance is useful both when we want to analyse tokens that contain hesitations and when we want to exclude such tokens. With some further programming effort the PD curves can also be used to tag the time extent of the potential hesitation, so that further analysis can disregard the uninteresting parts of the recording.

However, from the point of view of answering the main research questions of this thesis, the materials of this experiment are unideal. Even if we removed the tokens with hesitations and other extraneous movements from analysis, there would still be multiple problems.

The picture set used in this experiment was originally chosen to maximise the recognisability of the pictures (Snodgrass and Vanderwart 1980). However, there is ambiguity inherent in many of the pictures: the picture which is supposed to originally be an alligator can just as easily elicit the word 'crocodile', American and British English speakers have different names for many of the objects, and on top of this sometimes participants just fail to identify the object. The set was designed and tested almost 40 years ago with American speakers of English as the reference population. Whether the robustness of naming holds for other generations or speakers of other languages and from other cultures is debatable.

As it is, the Snodgrass picture set is *not* suitable for providing a phonetically diverse set of data with controlled syllable structure and balanced repetitions of a given phonetic onset. These are essential requirements for studying how the variation of utterance initial acoustic and articulatory timing is affected by phonetic content of the utterance.

In contrast, Experiment 2 has been specifically designed to provide such data. Analysing that experiment in the next chapter will enable us to answer the main research questions of this thesis.

Chapter 5

Experiment 2: Delayed Naming in UTI

The purpose of this experiment is to answer Research Question 1 and its more specific sub-questions about the timing relations of articulatory and acoustic onsets as laid out in Section 2.5.1). To recap, Question 1 and its sub-questions are:

Question 1: What is the relative timing (and absolute reaction time in relation to the go-signal) of tongue movement initiation (aka articulatory reaction time) and acoustic initiation (aka acoustic reaction time) in different phonetic contexts in a speech reaction time task, following instructions used by Rastle et al. (2005)?

Question 1a: Is articulatory reaction time affected by the acoustic duration of the onset consonant (OD) or by the acoustic duration of the utterance's rhyme?

Question 1b: Is acoustic reaction time affected by the acoustic duration of the onset consonant (OD) or by the acoustic duration of the utterance's rhyme?

Question 1c: Is the Articulatory to Acoustic onset Interval (AAI) affected by the acoustic duration of the onset consonant (OD) or by the acoustic duration of the utterance's rhyme?

As discussed in Section 2.5.1, in the hypothesis concerning Question 1a, we expect the articulatory reaction time measured from the tongue to correlate with Onset consonant's acoustic Duration (OD) and the duration of the utterance's rhyme. Furthermore, in the hypotheses concerning Questions 1b and 1c, we expect both the acoustic reaction time and the Articulatory to Acoustic onset Interval (AAI) to be inversely correlated with the OD and positively correlated with the duration of the utterance's rhyme.

In other words, the expectation is that the inverse correlation pattern in the data reported by Rastle et al. (2005) will be replicated in the acoustic reaction time results, and that its origin is the AAI. We also expect the silent articulation period – that is, the AAI, to be part of the word in articulatory terms.

The rhyme duration acts here as a proxy for the effect of speech rate. To avoid correlating part of a time interval with the whole, and to provide clear acoustic demarcation points for segmentation, the rhyme duration is measured as the duration of the acoustic interval from the *end* of the onset consonant to the beginning of the release burst of the final plosive in the /VC/ and /(C)(C)CVC/ words used as target words in this experiment.

The questions are answered in this chapter with statistical models to identify the stage where the effect of onset consonant duration on acoustic reaction times originates from. This is done by fractionating the acoustic reaction time into articulatory reaction time and the AAI.

The next section discusses the background of this experiment and aspects of experimental design in more detail. It is followed by sections on materials and methods, audio and articulatory analysis, results. The results section first goes through the statistical analysis to answer the research questions and then continues with an explorative part looking at the location of articulatory activation onset. The chapter ends with a section discussing the consequences of the results.

5.1 Introduction

This experiment measures articulatory and acoustic onsets using a delayed naming paradigm through an instruction on a computer display. In delayed naming the participant is told before a trial – usually by a computer display – what the target word is, then after the participant has indicated that they are ready to proceed, a go-signal is played after a randomly assigned delay. This is an effective way of isolating planning stages of speech production from the response delay. With an added instruction to remain at rest and not to prepare articulatorily for the target utterance, it is possible to examine the effect of phonetic content of the target word or utterance on the timing of speech initiation (Rastle et al. 2005, Kawamoto et al. 2008).

As discussed in Section 2.2.2, the acoustic study by Rastle et al. (2005) shows that the place and manner of the first consonant in a target affects acoustic reaction time. As discussed earlier (Section 2.2), Rastle et al. (2005) are very thorough in considering all phonotactically legal *consonantal* onsets of English. Rastle and colleagues also offer a careful analysis of factors such as voicing, manner and place of articulation, and vowel quality on acoustic reaction time, yet do not discuss in their article the strong inverse correlation of acoustic reaction time and OD evident in their data. (The data which was originally displayed in Figure 2.7 is repeated in Figure 5.1 for the reader's convenience). They also record and analyse only acoustic data of open syllables, leaving open the role of articulation in the timing of speech initiation and how the timing of initiation events relates to the timing of the whole utterance.

In contrast, (and also discussed in Section 2.2.2), the articulatory study by Kawamoto et al. (2008) shows for delayed naming with Rastle's revised instructions that the same effect is not present in articulatory reaction time of the lips and jaw, which they defined as "any noticeable changes in the lip and jaw position" extracted by manual examination of frontal lip videos. However, because they did not have access to articulatory data other than external recordings of lip and jaw movement, their phonetic materials were limited to only syllables

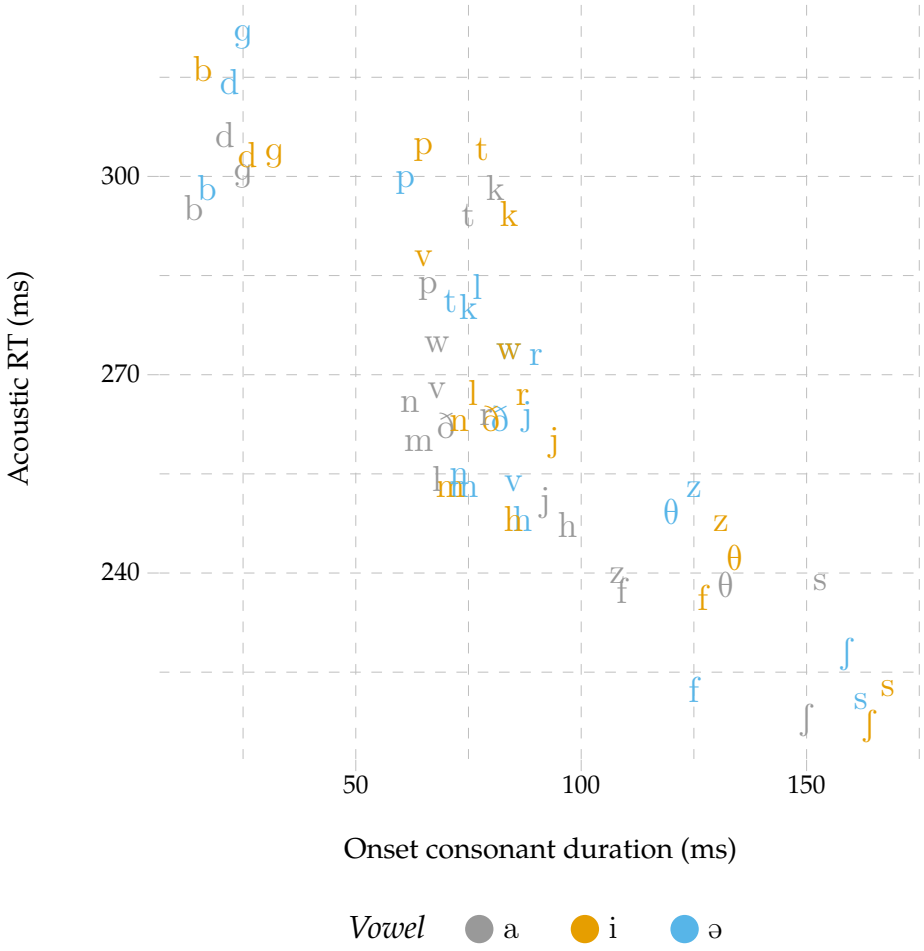

Figure 5.1: Correlation of consonantal onset duration with acoustic naming latency in delayed naming. This figure shows the data for /CV/ onsets reported by Rastle et al. (2005).

with bilabial and alveolar nasals and plosives (that is, /m,n,p,t/) as the onsets.

The *vowel* onsets, which were missing from the Rastle et al. (2005) experiment, were included in an articulatory study by Mooshammer et al. (2012) where they concluded that initiating the articulation of /VC/ words takes longer than initiating the articulation of /CV/ words. However, they recorded articulation with Electromagnetic Articulography (EMA), which – unlike Ultrasound Tongue Imaging (UTI) – does not capture movements of the back of the tongue nor tongue internal changes. Please refer to Chapter 6 for further analysis of the differences between UTI and EMA in analysing initiation timing.

In combination, these two studies suggest that the variation evident in Rastle and colleagues' data is actually variation in the time from articulatory onset to acoustic onset – the AAI. This hypothesis was put forward in 2.5 and is tested with this experiment.

This experiment is an expansion of the original experiment by Rastle et al. (2005). The original protocols needed to be adapted in several ways to be applicable to UTI and Standard Scottish English speakers. First, Rastle and colleagues' materials had a number of tokens (168 individual target syllables in total) from each participant which is not practical for UTI and a larger number of participants. In fact, they had each of the participants record the materials in six 45 minute sessions over two days. Limitations imposed by ultrasound meant that Experiment 2 here needed to cut the materials to a subset of those used in the earlier study. Otherwise, the recording sessions would either have been too long according to the design restrictions set out above or it would have only been possible to record each token only once from each speaker.

Second, the earlier study used phoneticians as participants and had them produce /CV/ syllables as specified with phonetic symbols. To facilitate recruiting participants who have no or little experience with phonetics, the materials were everyday words rather than syllables. This could have been avoided by using phoneticians as participants. While Rastle and colleagues managed to recruit five native speakers of Australian English as their participants, in general a sizeable sample of phoneticians with identical language backgrounds is difficult to recruit. Furthermore, use of everyday words avoids the confound of mixing nonsense words with lexical words. Thus, all of the participants were Standard Scottish English monolingual speakers (self-identified and verified by the author) and were recruited from among the staff and students of Queen Margaret University.

Design of the UTI experiments

Participant comfort posed a practical limit on the length of a single continuous recording session in UTI. With very few exceptions, participants can endure

wearing the stabilisation headset for at least 20 minutes without breaks. Thus, a recording session of about 1.5 hours with a 10-minute-setup of the headset, two 20-minute-long recording blocks, and a break of 30 minutes or slightly more in between form the basis for designing the UTI experiments.

Recording one short token – for example, a word – takes a bit under 15 seconds including the saving time required by the ultrasound data and the full RGB videos. As a result, recording of 90 tokens takes about 20 minutes. This means that the experiments can have at most 180 tokens when divided into two blocks. The helmet needs to be taken off for the rest between the blocks and this makes it impossible to guarantee that the probe is in the same place during the two blocks.

Taking the ultrasound helmet off would pose a problem for analysis methods which rely on being able to compare anatomically matched data, since there could be no guarantee that the probe would rest in exactly the same location for every recording batch. However, when using holistic methods like Pixel Difference (PD) for analysis, this problem can be avoided, as the analysis of the whole image does not require such precise positioning. This approach offers greater freedom in designing the data sets. Not only is it possible to break recording of large numbers of tokens into batches as described above, but this also makes it possible to add to existing data sets by recruiting the same participant at a later date to record more data.

In Experiment 1, the participants were recorded with a bite plate in their mouth to measure the orientation of the UTI probe in relation to the occlusal plane. In Experiment 2, and the UTI part of Experiment 3, the bite plate procedure was repeated at the beginning and end of each block with a water swallow to provide a palate trace. These procedures were performed to provide for the future possibility of using tongue surface contour extraction in further analysis of the experiments, but that work falls outside of this thesis.

5.2 Materials and methods

5.2.1 Participants

Four monolingual speakers of Standard Scottish English (dialect verified by self-reports and by the author) were recorded in this experiment. All participants had either normal or corrected-to-normal vision and no known hearing or speech problems. The participants were assigned identifiers P1-P4. Their background information and the time that each participants' recordings spanned is summarised in Table 5.1. P2 recorded only part of the experiment. She also had very long reaction times, and thus her data is included only in the initial analysis below.

Table 5.1: Background information and time span of recordings for participants of Experiment 2.

Speaker	Gender	Year born	Age	Phonetical training	Span
P1	Male	1989	26/27	Yes	11 months
P2	Female	1983	32	Yes	3 days
P3	Female	1988	27/28	No	9 months
P4	Female	1988	28	Yes	1 month

According to the terms of the ethical approval obtained from Queen Margaret University. Prior to recording, the whole recording procedure was explained to the participants. They were also given the opportunity to withdraw from the experiment at any time without needing to state a reason. The purpose of the experiment was not disclosed initially to the participants, but they were given the opportunity to be debriefed after completing their recordings (or after having decided to withdraw from the experiment).

Informed consent was obtained from all participants before recording and all participants filled a background questionnaire during the first recording session. Participants were reimbursed £10 for each recording session. The data was anonymised and the contents of the background information forms were transcribed and their contents also anonymised. The original filled forms are retained in secure storage at Queen Margaret University.

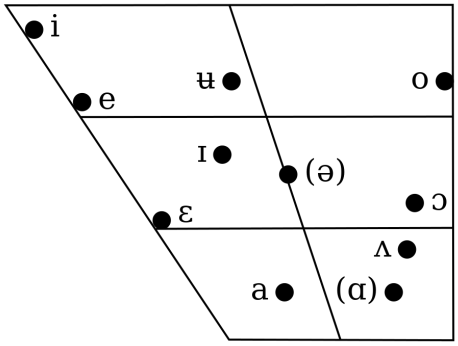

Figure 5.2: Monophthongs of Standard Scottish English (Scobbie et al. 2007).

5.2.2 Stimuli

Rastle et al. (2005) used three vowels: [a, ə, i]. Two of these are readily available in Standard Scottish English in /CVC/ lexical words, but [ə] is not. Using only three vowels is a practical necessity as the number of target words needs to be limited and the focus of the experiment is on the effect of the onset segment – not the quality of the syllable nucleus. There are two criteria for optimising the choice of three vowels if freely sampling from the whole vowel space: Either, take three extreme vowels – for example, [a, i, u] – thus reducing the quadrilateral space to a vowel triangle, while maximising the covered area, or – and this is the approach adopted by Rastle and colleagues – take two of the most extreme vowels [a, i] and the vowel at the centre of the system [ə] further reducing the area of vowel space covered by the triangle, while maximising vowel variation *and* variation in the articulatory effort needed to produce the vowels.

The first approach covers the whole vowel space but has only extreme articulations, which means that, if we were interested in gradient effects across variables making up the vowel space, these can not be studied with this approach. The second approach corrects for this problem, but does it with the trade off of reducing the quadrilateral vowel space to a compressed triangle – one covering less than half of the vowel space. The first approach has been used in this experiment because there are no stressed syllables with a [ə] nucleus in the phonological system of Scottish English and it was considered desirable to

avoid adding syllable stress variation as a confound in the materials. [u] could not be used as one of the vowels, since the phonological English /u/ is realised in Scottish English as the more centralised [ʉ]. Using [o] was further ruled out by the poor availability of suitable words filling the word selection criteria below. The vowels were thus chosen to be [a, i, ɔ].

As for the selection of onset consonants, a first simplification was to vary the vowel context only with simple consonantal onsets and have complex onsets in only one of the vowel contexts. After this, the choice was guided by analysing the acoustic reaction times and onset durations of the data reported by Rastle and colleagues. Looking at Figure 5.1, which plots the acoustic duration of the initial consonant against acoustic reaction time in the data reported by Rastle et al. (2005), it can be seen that certain sounds cluster together and that most of the sounds pattern with a linear inverse relation between OD and acoustic reaction time. Unvoiced plosives form a clear exception to the general pattern.

As seen in Figure 5.1, onset *consonant* duration plays an interesting role in the patterning of delayed naming data. To make it possible to get an adequate number of repetitions for each word, it was necessary to limit the number of onset consonants. Thus, to obtain a balanced set of different manners of articulation and different OD and acoustic reaction time combinations, the following onset consonants were chosen for this experiment (roughly in order of increasing acoustic reaction time): /s, ʃ, f, n, m, h, l, r, w, p, t, k, b, d, g/. Following the example of Mooshammer et al. (2012), a set of vowel onset words was included in the materials. The three vowels chose above – /a, i, ɔ/ – were used as the onsets and also served as the nuclei of the consonant onset words.

The words were chosen from a pronunciation lexicon of Standard Scottish English generated automatically with Unisyn (Fitt 2014). The dictionary was searched for word triads, which had the same onset consonant, one each of the vowels /a,i,ɔ/. A further requirement was that the final consonant was a voiceless stop – that is, one of /k,p,t/. This choice was made to produce a clear acoustic marker – the release burst – that could be used to calculate the acoustic duration of the rhyme of the word, in those productions where the participant

released that final stop.

Finally, to avoid confounding the results with a possible lexical frequency effect (Kawamoto et al. 2008), the set used in this study was selected by maximising the lexical frequency of the target words as provided in the Unisyn (Fitt 2014). With the restrictions listed above it was not possible to always select words with very high lexical frequencies and instead the selection was driven by the aim to avoid words with very low lexical frequencies. Nevertheless, two possibly problematic words were chosen as part of the set: 'DAT' and 'Nat'. However, as discovered by Jescheniak and Levelt (1994) (cited by Kittredge et al. (1994)), there is no lexical frequency effect in delayed naming of pictures. This makes it unlikely that there will be one in delayed naming when the target word is represented in written form. Systematic analysis of the effect would require the data set to be designed in a different way, and thus is outwith the scope of this thesis. In the present context, any residual frequency effect will be included in the random effect of word in the linear mixed models fitted to the data.

To summarise, the 58 chosen words were of the four types /CCCVC/, /CCVC/, /CVC/, and /VC/. The chosen target words are listed in Table 5.2 according to their phonetic onsets and syllable nuclei.

Table 5.2: Target words used Experiment 2 listed by onset complexity and initial phoneme.

/VC/	/a/	/i/	/ɔ/
#	at	eat	ought

/CVC/	/s/	/ʃ/	/f/	/n/	/m/	/h/	/l/	/r/	/w/	/p/	/t/	/k/	/b/	/d/	/g/
/a/	sat	shack	fat	Nat	map	hat	lack	rat	whack	pack	tap	cat	back	DAT	gap
/i/	seat	sheet	feet	neat	meet	heat	leap	reap	wheat	Pete	teak	keep	beat	deep	geek
/ɔ/	sought	shop	fought	not	mock	hot	lot	rock	what	pop	talk	caught	bought	dot	got

/CCVC/	/s/	/ʃ/	/f/	/n/	/m/	/h/	/l/	/r/	/w/	/p/	/t/	/k/	/b/	/d/	/g/
/a/			flat							prat		crap	black	drat	
/i/	steep	shriek									treat				greek

/CCCVC/	/s/	/ʃ/	/f/	/n/	/m/	/h/	/l/	/r/	/w/	/p/	/t/	/k/	/b/	/d/	/g/
/i/	street														

5.2.3 Procedure

In the experiment that this experiment replicates, Rastle et al. (2005) recorded each word six times from each participant. We decided to base this experiments recordings on two-day sessions, during which each word would be recorded six times over the course of two days. In an effort to guarantee statistical power and to provide a future possibility of analysing the data for changes in speakers productions over a longer time period, each speaker was invited to repeat the two-day session a total of three times.

Since there are 58 words in the target word list, recording six repetitions of each means recording a total of 348 separate tokens. Repeat this three times, and we have a goal of recording 1044 tokens from each participant. As we see in Section 5.5, the goal was reached with three participants.

However, for logistical reasons it proved difficult to get an even spacing of the recording sessions for each participant. This can be seen in the recording periods reported in Table 5.1. The sessions were spread over several months for participants P1 and P3, while P2 was recorded over a period of three days and P4 over a period of one month.

Recording 348 tokens takes about 80 minutes. To make this feasible – given the 20 minute limit on a having the ultrasound headset on – the session was split into four batches of about 20 minutes each. The batches were further split over two days – two batches per day – with 30 minute break between the batches during a given day. So, each recording session lasted an hour and a half on two days, with the six sessions spread over a longer period of time (Table 5.1).

During each two-day session the words were produced six times in a block-wise randomised order. The randomisation of the blocks was implemented in R Core Team (2013) so that it guaranteed that words were not repeated across block boundaries (which would be the result of the previous block ending with the word the next block begins with). During every recording day, the headset was taken off between the 20-minute batches to give the participant a rest and it had to be adjusted to fit the participants head anew for every day of recordings.

Each trial consisted of the following sequence: The participant read the

next target word from the computer screen. When the participant felt that they were ready to produce the word, they indicated so by pressing a button on a keyboard. The key press activated the sound and ultrasound recording. The experimental software automatically initiated the ultrasound recording about 0.5 s after the sound recording began providing an adequate window to examine the stability of the participant's articulation before the go-signal was given. After a random delay, which was uniformly distributed between 1200 ms and 1800 ms, a go-signal – a 50 ms long 1000 Hz pure tone – was played via the computer's loudspeakers.

It was emphasised to the participant that it was important to keep their mouth (lips, tongue, and jaw) at rest until they heard the go-signal. It should be noted that no instructions were given about what kind of rest position the participant should employ. The only instruction was to "remain still, at rest". This was an intentional choice to facilitate studying the rest positions employed by the participants. The relevant analysis remains future work.

Before the experiment began, the participants were instructed to wait for the go-signal in each trial and after they heard it to "read the word out loud as fast and as accurately as possible keeping in mind that this is a speeded trial." As mentioned above, care was taken not to discuss the purpose of the experiment with the participants before they had completed all of the recordings. Therefore, the participants did not have certain knowledge of why it was important to stay at rest at the beginning of each trial, but it is always possible that they – especially those with phonetic training – may have guessed the reason.

The recording of most trials was manually terminated by the experimenter (the author) when they had heard the participant respond. However, some of participant P3's data was cut short by premature manual termination of the recording. After this was noticed, the procedure was changed to a automatic recording length of 4 seconds. P3 was also re-recruited to record enough data to replace the tokens that had been cut short. More precise details are given in Section 5.5.

5.2.4 Audio and UTI recordings

This experiment was run with the same setup as Experiment 1: synchronised ultrasound, lip imaging and sound recording were controlled with Articulate Assistant Advanced (AAA) software (Articulate Instruments Ltd 2012). The participants were fitted with a purpose-built headset to ensure stabilisation of the ultrasound probe (Articulate Instruments Ltd 2008). Attached to the helmet was a small Audio Technica AT803b microphone for high-quality acoustic recordings. Ultrasound recordings were obtained at a frame rate of 120 frames per second – with the exception of the first two batches for participant P1, which were captured at a frame rate of 83 frames per second due to a technical error. The high speed SonixRP system at Queen Margaret University (Wrench and Scobbie 2011) was used for all of the recordings.

An NTSC micro-camera was used to capture recordings of the speakers' lips. Video was captured at 30 fps, de-interlaced, and can be analysed at 60 fps. In this thesis the videos were only used to control for speech and/or articulation artefacts – such as false/early starts – in manual analysis of the recordings.

5.3 Audio analysis

The analysis of audio data works in steps. First, the stimulus onsets for each token were detected automatically with a Python script which was written for the purpose using NumPy and SciPy libraries (Python Software Foundation 2017). The script is described in detail below in the next section and included in Appendix B. Second, a raw acoustic segmentation was obtained with forced alignment with the FAVE forced aligner (Rosenfelder et al. 2011). Third, the author manually corrected the raw alignment results from FAVE in Praat (Boersma and Weenink 2010).

Rosenfelder et al. (2011) is a forced alignment system that provides a rough phonetic alignment of an audio sample. In order to do so, FAVE needs the audio sample and a .csv file (a file in the comma separated values format) listing the time intervals where speech occurs, together with an orthographic

transcription of the speech within each interval. FAVE then uses a pronunciation dictionary to map the orthographic words to possible phonetic representations. If a given word does not appear in FAVE's dictionary a phonetic mapping for it can be provided in an auxiliary text file. After the phonetic mapping is done, FAVE searches for the best alignment of phone boundaries using a Hidden Markov Model (HMM) of speech to decide where a segment begins and where it transitions to the next one. The segmentation obtained from FAVE is limited in precision because the Hidden Markov Model has to use a fairly long time window as the basic processing unit in order to keep the computational load manageable. However, the segmentation is accurate enough to make hand-correcting the boundaries more efficient than performing manual segmentation from scratch.

5.3.1 Automated detection of stimulus onset

The first stage of the audio analysis was to identify the 50 ms long 1 kHz beeps so that their onsets could be used as reference times for both articulatory and acoustic reaction times and so that they could be excluded from the intervals given to FAVE for forced alignment. For this purpose a Python script was written. It also prepares token metadata (mainly target words, participant IDs) for FAVE and later analysis stages.

The go-signal detector works in the following stages. First, it bandpass filters the audio recording on a narrow pass band centered at 1 kHz. Second, it finds a rough estimate of the beginning of the beep from the intensity contour of the bandpassed signal. Third, it refines the estimate by examining a neighbourhood of the rough estimate. This is accomplished by exploiting the fact that the beep is a sinusoid which starts with the signal level rising up, followed by it dropping down. Since physical acoustic signals are always damped to some degree, the initial rise is weaker than the second dip. Exploiting this knowledge, the script locates the first strong downward movement of the acoustic signal in the examined neighbourhood. Finally, it tracks from that oscillation to the next zero crossing, which will be the end of the first full wave of the sinusoid.

The final estimate is then calculated by deducting 1 ms (the period of a 1 kHz sinusoid) from that zero crossing. This results in sub-millisecond accuracy in detection of the signal onset.

5.3.2 Manually corrected forced alignment of the audio signal

FAVE was run locally on a MacBookPro laptop (Rosenfelder et al. 2011) to produce the raw acoustic segmentation. The pre-processing scripts described above provided FAVE with a beginning estimate of where to look for a given word. Some of the target words – and especially their Scottish English pronunciations – were not in FAVE's pronunciation dictionary. Supplemental pronunciation rules were provided for FAVE in an auxiliary input file. Based on this input FAVE produced the raw alignment as a Praat textgrid. To produce the actual alignment, the segment boundaries of the raw alignment results from FAVE were manually corrected in Praat (Boersma and Weenink 2010) and re-mapped the segment's phonetic identities based on Scottish English pronunciation of the target words. The last step was necessary because the pronunciation model in FAVE is based on American English dialects and often produced unrealistic vowel qualities in this data.

Below is the list of rules that were followed when correcting the raw alignment results from FAVE. The rules were derived from those set out by Hewlett and Beck (2006) and Turk et al. (2006). Some adaptations were necessary based on the actual productions of the participants.

First, some general rules:

- Pre-aspirations were marked as separate intervals, but included in the onset consonant duration.
- In case of clear final consonant releases the onset of the release burst was marked to provide an end-of-word boundary.
- In case the vowel did not properly end (recording was cut short), the final consonant segment was removed from the transcription to mark this.
- In case the first consonant did not properly end (recording cut even

shorter), the vowel segment was removed from the transcription.

Second, onsets and offsets for C_1 and V:

- Vowels: Onset was defined by offset of previous consonant, if present, or by the onset of phonation including possible onset burst.
  Offset was defined as either the first significant dip in waveform amplitude at the end of the vowel-like part of the signal and/or the end of regular phonation and/or the loss of vowel-like formant structure.
- [f, h, s, ʃ]: Onset was defined as the onset of frication and offset as the onset of phonation.
- [m, n]: Onset was defined as the onset of phonation as evident on the waveform.
  Offset was defined as either a dip in the waveform amplitude, or a shift formants – either formant frequencies or, for example, a previously suppressed F4 gaining energy. Often there was a clearly observable transient visible on the spectrogram indicating separation of tongue from the palate. When present, this was used as the segmentation boundary.
- [l]: Onset was defined as onset of phonation, excluding any frication noise that might be present before phonation onset. The basis of this decision is that it is not possible to clearly distinguish between frication following opening 'smacks' and frication due to constrictions tightening up.
  Offset was defined as either a dip in the waveform amplitude, or a shift formants – either formant frequencies or, for example, a previously suppressed F4 gaining energy. When other criteria were absent, but a gradually rising F2 was present, the offset was set at the middle of the rise as this seemed to coincide with the dip in amplitude when that was present.
- [r] Onset was defined as the onset of phonation and offset as the beginning of a shift in formant frequencies or the middle of a dip in waveform amplitude.
- [w]: Onset was defined as either the onset of frication – if there was preaspiration present – or the onset of phonation. In analysis, the length of

the sound was defined as sum of any pre-aspiration and following voiced portion.
Offset was defined as a dip in waveform amplitude and/or change in formant frequencies.

- [b,d,g]: Onset was defined as either the onset of the release burst or onset of phonation if the sound was fully phonated.
  Offset was defined as the onset of phonation of the following vowel or consonant.
  In further analysis, the pre-phonation and burst periods were grouped together when calculating the acoustic duration of the consonant.
- [p,t,k]: Onset was defined as either the onset of the release burst or the point where the preceding fricative's noise had significantly reduced.
  Regardless of whether [p,t,k] were part of a cluster or word initial, the release burst and aspiration were segmented as one separate interval, which was then added to the total duration of the consonant.
  Offset was defined as the onset of phonation of the following vowel or consonant.
  In further analysis these pre-phonation and burst were grouped together when calculating the acoustic duration of the consonant.

5.4 Articulatory analysis

5.4.1 Manual labelling of articulatory onset

Articulatory onsets were labelled in a subset of the data on ultrasound videos in AAA (Articulate Instruments Ltd 2012) to provide a baseline for validating the automated onset detection methods described in Chapter 3. The subset consisted of first two days of recordings from participants P1, P2 and P3. This meant that from P1 and P3 a third (1/3) of the full data set was labelled, but since P2 only recorded the first two days out of six, all of her data was labelled. In overall terms, this means that out of the pooled data of P1-P4 given the different

amounts of data recorded for each participant, about 27 % of the recorded tokens were manually labelled.

5.4.2 Localised tongue movement onsets

In addition to the manual labelling described above, all of the data in this experiment was analysed using Scanline Based Pixel Difference (SBPD) in the fully automated mode. The basic analysis sequence is described in more detail in Section 3.5. Following the calculation of scanline-based onsets each token's articulatory onset was operationalised as the median of scanline-based onsets as detailed in Section 3.6.

5.5 Results

The recordings of this experiment form a large UTI corpus. Raw sample sizes for the individual participants were P1: 1074, P2: 347, P3: 1254, and P4: 1044. These sizes were calculated after missed trials were removed but before various thresholding operations which are dependent on the analysis method.

A good deal of the recordings that were cut short by the experimenter ending the recording too soon. Unfortunately, this was not immediately evident and caused significant data loss with especially the trials with long intervals from trial onset to go-signal with P3. Once the situation was noticed, the recordings were changed from manual cut off to automated cut off at 4 seconds. Four additional 20 minute session were recorded with P3 using go-signal delays fitted to compensate the lost trials. This rectified the data loss, by balancing the sampled delay distributions.

The most common reason for excluding a token from analysis was that the recording was cut too short (before vowel onset) to provide reliable data, but some tokens were also removed as mispronunciations during the manual correction of audio segmentation. All of these were labelled with an acoustic onset set within the beep, so that they were later automatically dropped from further analysis when tokens with too early onsets were removed.

5.5.1 Statistical models

To answer research questions 1a-1c, the data was fitted with linear mixed effects models to explain variation in articulatory reaction time, acoustic response time and in the AAI. The results for acoustic reaction time and AAI show that both variables are negatively correlated with OD and positively correlated with rhyme duration. In contrast, for articulatory reaction time the results show that it is correlated with neither OD nor rhyme duration.

Before statistical analysis the data was thresholded with the lower bound of 28 ms for the SBPD onset (as set out in Section 3.3.2 based on Chiu and Gick 2014), the acoustic reaction time between 58 ms and 500 ms, and AAI with a lower bound of 0 ms. Participant P2 was removed from the data set because she produced very few tokens with acoustic response times under 500 ms. Tokens where the audio recording was incomplete and where no clear release of the final consonant was evident (which renders it impossible to measure rhyme duration) were also removed from the data set. Finally, the data set was restricted to include only /VC/ and /CVC/ words, since including complex onsets would have unbalanced the data set with very few tokens coming from the complex classes. The data set used in the models and graphs in this section consists of 1386 tokens – 439 from P1, 672 from P3, and 275 from P4. The differences are mainly due to different proportions of tokens with a detectable final release of the final consonant in each speaker's data.

The data was analysed in R (R Core Team 2013) by iteratively fitting linear models (step-up process Baayen 2008) with version 1.1-21.9000 of the `lme4` package (Bates et al. 2015) to explain the variation in articulatory reaction time, acoustic reaction time, and AAI. Step-up comparisons were performed with the built-in `anova` function of R.

Articulatory reaction time

When the process was carried out on SBPD based articulatory onsets, the first model included only random effects and was (in R formula notation):

$$ArtRT \sim (1|id) + (1|word) \tag{5.1}$$

where $(1|id)$ is the random effect for the participant and $(1|word)$ correspondingly for the word. Including trial number (representing how far in a given recording batch the sample was) in Model 5.1 we have:

$$ArtRT \sim trial + (1|id) + (1|word). \tag{5.2}$$

Testing with ANOVA whether the second model has improved the fit yields a statistically significant result: P-value ≈ 0.017 evaluated on the χ^2 distribution with one degree of freedom.

Summary of the final model – fitted after removing outliers by culling data points with an absolute value of the normalised residual in excess of 2.5 – is given in Table 5.3. In the table we see that coefficient of the only fixed effect is trial: -0.076, meaning that the participants got on average slightly faster over the course of a batch of recordings.

All other models for predicting articulatory reaction time failed to reach significance. Together with the model discussed next this indicates that most of the variation in acoustic reaction time should be attributed to AAI apart from the trial effect, and that articulatory reaction time is mainly a noisy constant from the point of view of phonetic variables.

Acoustic reaction time

The iteration process identified a statistically significant model that successfully predicts acoustic reaction time ($AcRT$ in the formulas). The first model included only random effects and was (in R formula notation):

$$AcRT \sim (1|id) + (1|word) \tag{5.3}$$

where $(1|id)$ is the random effect for the participant and $(1|word)$ correspondingly for the word. Including the acoustic duration of the onset consonant (OD)

in Model 5.3 we have:

$$AcRT \sim OD + (1|id) + (1|word). \tag{5.4}$$

Testing with ANOVA whether the second model has improved the fit yields a statistically significant result: P-value $< 2.2 \times 10^{-16}$ evaluated on the χ^2 distribution with one degree of freedom. Including rhyme duration ($RhymeDur$) in Model 5.4 we have:

$$AcRT \sim OD + RhymeDur + (1|id) + (1|word). \tag{5.5}$$

Table 5.3: Summary of the final articulatory reaction time mixed effects model.

```
Linear mixed model fit by REML ['lmerMod']
Formula: art_RT ~ trial + (1 | id) + (1 | word)
   Data: subset(RTs, abs(scale(resid(RTs.lmer_art))) < 2.5)

REML criterion at convergence: 13149.7

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.2705 -0.6545 -0.0484  0.5917  3.5650

Random effects:
 Groups   Name        Variance Std.Dev.
 word     (Intercept)   43.45   6.591
 id       (Intercept)  186.12  13.643
 Residual             1040.28  32.253
Number of obs: 1339, groups:  word, 48; id, 3

Fixed effects:
             Estimate Std. Error t value
(Intercept) 120.51666    8.12581  14.831
trial        -0.07609    0.01759  -4.325

Correlation of Fixed Effects:
      (Intr)
trial -0.182
```

Testing with ANOVA whether the third model has improved the fit yields a statistically significant result: P-value $\approx 7.616 \times 10^{-9}$ evaluated on the χ^2 distribution with one degree of freedom. Including trial number ($trial$) in Model 5.5 we have:

$$AcRT \sim OD + RhymeDur + trial + (1|id) + (1|word). \tag{5.6}$$

Testing whether the third model has improved the fit yields a marginally statistically significant result: P-value ≈ 0.04548 evaluated on the χ^2 distribution with one degree of freedom. No further statistically significant effects were identified.

Summary of the final model – fitted after removing outliers by culling data points with an absolute value of the normalised residual in excess of 2.5 – is given in Table 5.4. In the table we see that coefficients of the fixed effects are OD: -0.42, rhyme duration: 0.20, and trial: 0.06. It is worth noting that the trial effect has the opposite sign when compared with the trial effect on articulatory reaction time, meaning that acoustic reaction times of the participants got relatively longer towards the end of a recording session.

Articulatory to Acoustic onset Interval

Similarly to acoustic reaction time, the iteration process identified a statistically significant model that successfully predicts AAI. The first model was:

$$AAI \sim (1|id) + (1|word) \tag{5.7}$$

where $(1|id)$ is the random effect for the participant and $(1|word)$ correspondingly for the word. Including the acoustic duration of the onset consonant (OD) in Model 5.7 we have:

$$AAI \sim OD + (1|id) + (1|word). \tag{5.8}$$

Table 5.4: Summary of the final acoustic reaction time mixed effects model.

```
Linear mixed model fit by REML ['lmerMod']
Formula: ac_RT ~ OD + rhyme_dur + trial + (1 | id) + (1 | word)
   Data: subset(RTs, abs(scale(resid(RTs.lmer_ac3))) < 2.5)

REML criterion at convergence: 13878.4

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.56740 -0.74671 -0.06237  0.63586  2.95382

Random effects:
 Groups   Name        Variance Std.Dev.
 word     (Intercept)  216.9   14.73
 id       (Intercept) 2821.5   53.12
 Residual             1447.9   38.05
Number of obs: 1362, groups:  word, 48; id, 3

Fixed effects:
             Estimate Std. Error t value
(Intercept) 268.85914   31.96871   8.410
OD           -0.42418    0.03897 -10.884
rhyme_dur     0.20489    0.03272   6.262
trial         0.06291    0.02076   3.030

Correlation of Fixed Effects:
          (Intr)   OD   rhym_d
OD        -0.070
rhyme_dur -0.257 -0.001
trial     -0.035 -0.018 -0.070
```

Testing whether the second model has improved the fit yields a statistically significant result: P-value $\approx 1.057 \times 10^{-13}$ evaluated on the χ^2 distribution with one degree of freedom. Including rhyme duration ($RhymeDur$) in Model 5.8 we have:

$$AAI \sim OD + RhymeDur + (1|id) + (1|word). \tag{5.9}$$

Testing whether the second model has improved the fit yields a statistically significant result: P-value $\approx 4.252 \times 10^{-11}$ evaluated on the χ^2 distribution with one degree of freedom. Including trial number ($trial$) in Model 5.9 we have:

$$AAI \sim OD + RhymeDur + trial + (1|id) + (1|word). \tag{5.10}$$

Testing whether the second model has improved the fit yields a statistically significant result: P-value ≈ 0.0003768 evaluated on the χ^2 distribution with one degree of freedom. No further statistically significant effects were identified.

Summary of the final model – fitted again after removing outliers by culling data points with an absolute value of the normalised residual in excess of 2.5 – is given in Table 5.5. In the table we see that the coefficients of the fixed effects are OD: -0.40, rhyme duration: 0.31, and trial: 0.12. Now we see that the AAI has a trial effect of roughly twice the size of the same effect on articulatory reaction time and of opposite sign. This means that the trial effect on acoustic reaction time is the result of the effect on AAI dominating the one on articulatory reaction time.

Figures 5.3 and 5.4 show that the inverse correlation of both acoustic reaction time and AAI as predicted by the above statistical models are clear enough that they are detectable in these scatter plots. Furthermore, each of the participants repeats the correlation patterns independently. Figure 5.5 shows that the case for AAI correlating with rhyme duration is not as clear for P1, but P3 and P4 do show a clear positive correlation in their distributions.

Table 5.5: Summary of the final Articulatory to Acoustic onset Interval mixed effects model.

```
Linear mixed model fit by REML ['lmerMod']
Formula: AAI ~ OD + rhyme_dur + trial + (1 | id) + (1 | word)
   Data: subset(RTs, abs(scale(resid(RTs.lmer_AAI_3))) < 2.5)

REML criterion at convergence: 14034.7

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.94704 -0.62879 -0.02521  0.62151  2.86973

Random effects:
 Groups   Name        Variance Std.Dev.
 word     (Intercept)  500.6   22.38
 id       (Intercept) 2206.3   46.97
 Residual             1818.4   42.64
Number of obs: 1345, groups:  word, 48; id, 3

Fixed effects:
             Estimate Std. Error t value
(Intercept) 120.06059   29.12408   4.122
OD           -0.39581    0.04532  -8.734
rhyme_dur     0.30577    0.03776   8.098
trial         0.12176    0.02329   5.228

Correlation of Fixed Effects:
          (Intr)   OD   rhym_d
OD        -0.087
rhyme_dur -0.326 -0.004
trial     -0.042 -0.030 -0.067
```

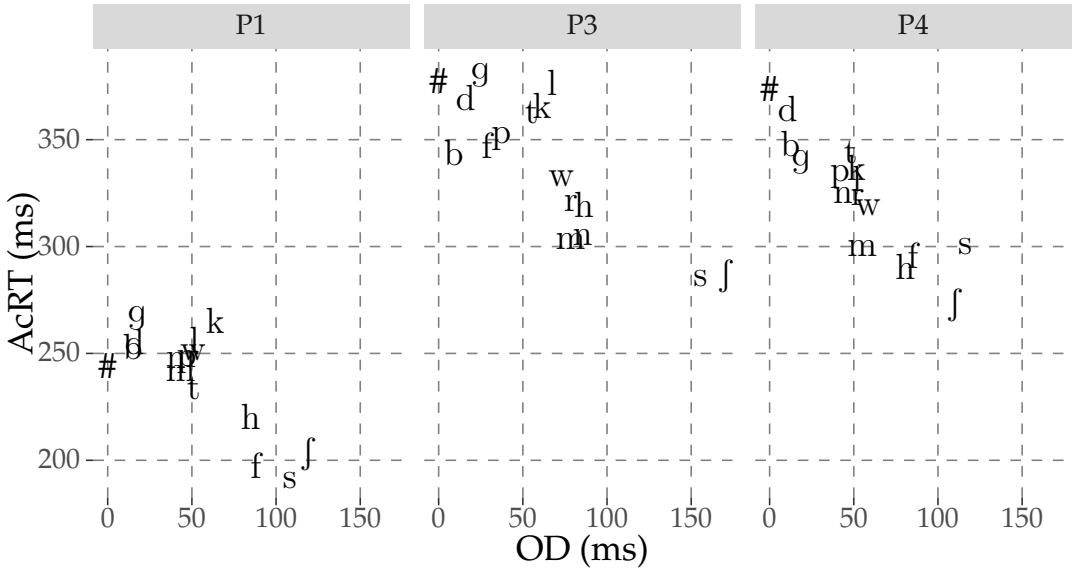

Figure 5.3: Acoustic reaction time as a function of OD for each modelled participant (P1, P3, and P4) medianised across tokens with the same onset consonant. # marks onsetless tokens, that is, /VC/ words.

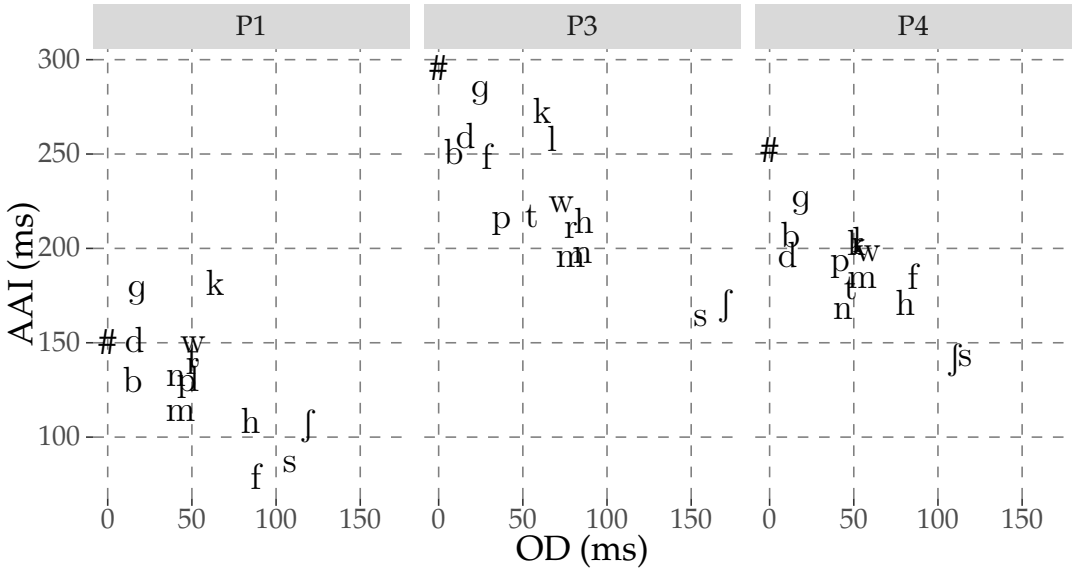

Figure 5.4: AAI as a function of OD for each modelled participant (P1, P3, and P4) medianised across tokens with the same onset consonant. # marks onsetless tokens, that is, /VC/ words.

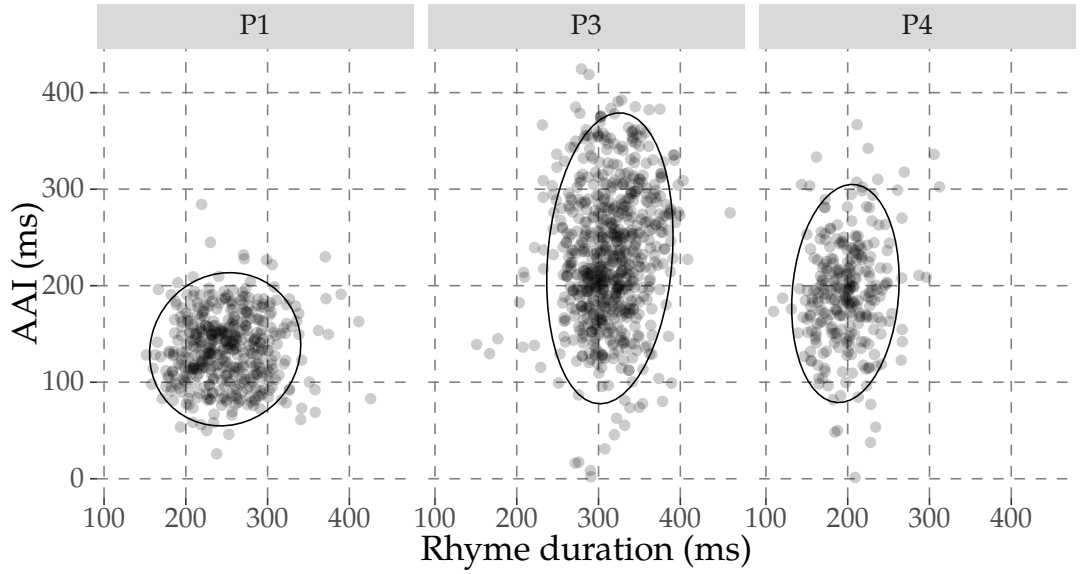

Figure 5.5: AAI as a function of rhyme duration (from the onset consonant's acoustic offset to the release burst of the final consonant) for each modelled participant (P1, P3, and P4). The graphs also include *t*-distribution-based 95% confidence ellipses.

5.5.2 Local movement onsets

Figures 5.6 - 5.9 show local onsets for each speaker. The graphs were produced with the scanline-based method described in Section 3.5.

The patterns are individual – compare, for example, the vowel onset panels (marked with #) of P1 and P3. Within individuals there is systematicity in that onsets with the same or similar place of articulation produce similar patterns suggesting that they employ similar motor strategies.

The most anterior region (corresponding to scanlines above number 50) contains the mandible shadow in the ultrasound data, accounting for the frequent late onsets in this region. Even so, in some contexts some of the participants have early movement in this region, implying that they start by moving their mandible. However, the general trend is for the first movement to be in the posterior region of the scanned area – usually scanlines below 40, or even as far back as scanlines under 20 for, for example, P1 in vowel and /ɹ/ onsets.

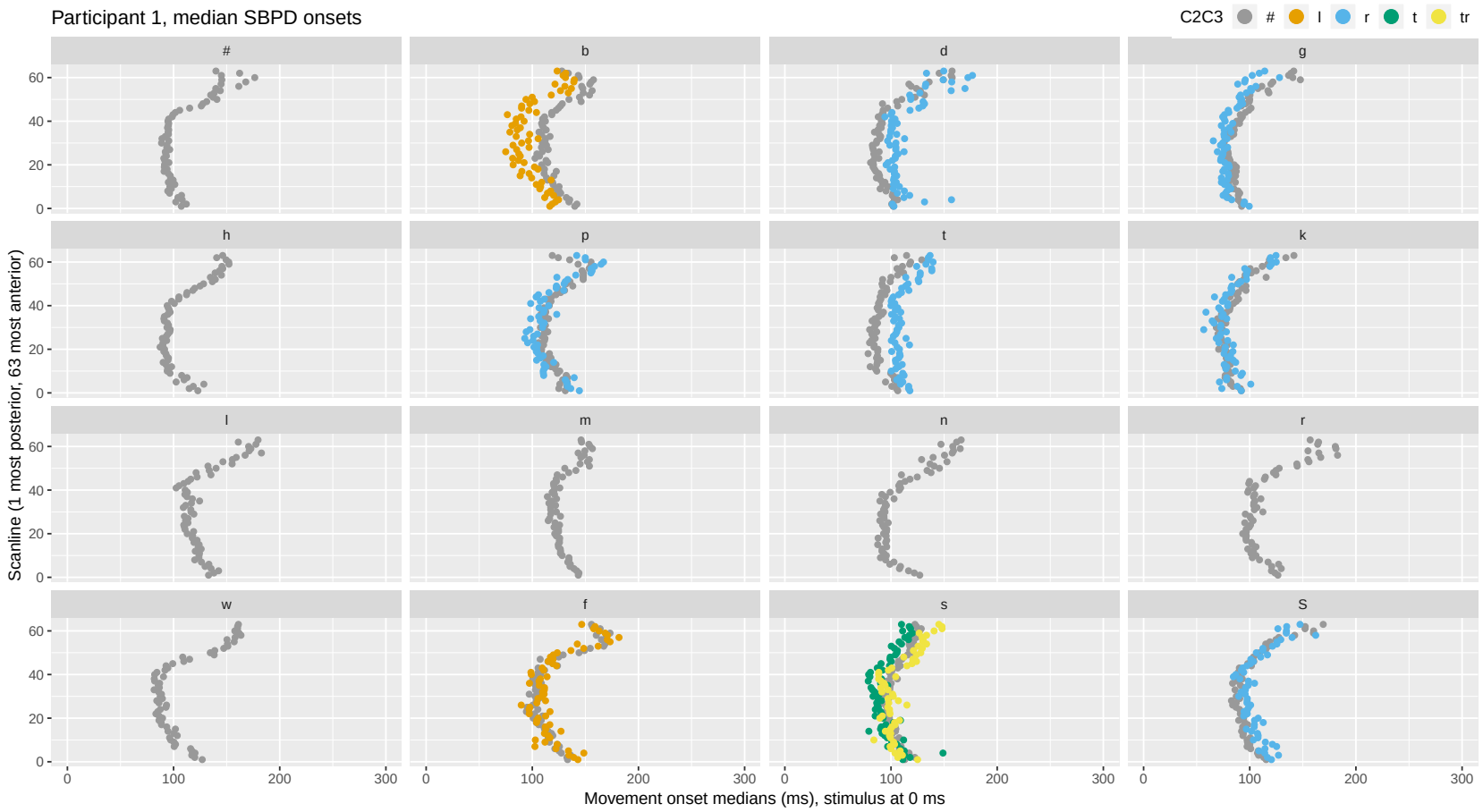

Figure 5.6: Medians of localised movement onset latencies for participant P1. Y-axis is time and x-axis position from back (1st scanline) to front (63rd scanline) conditioned by initial consonant (# marks no onset, that is, a /VC/ word). Grey dots correspond to /CVC/ utterances and coloured ones to utterances with complex onsets.

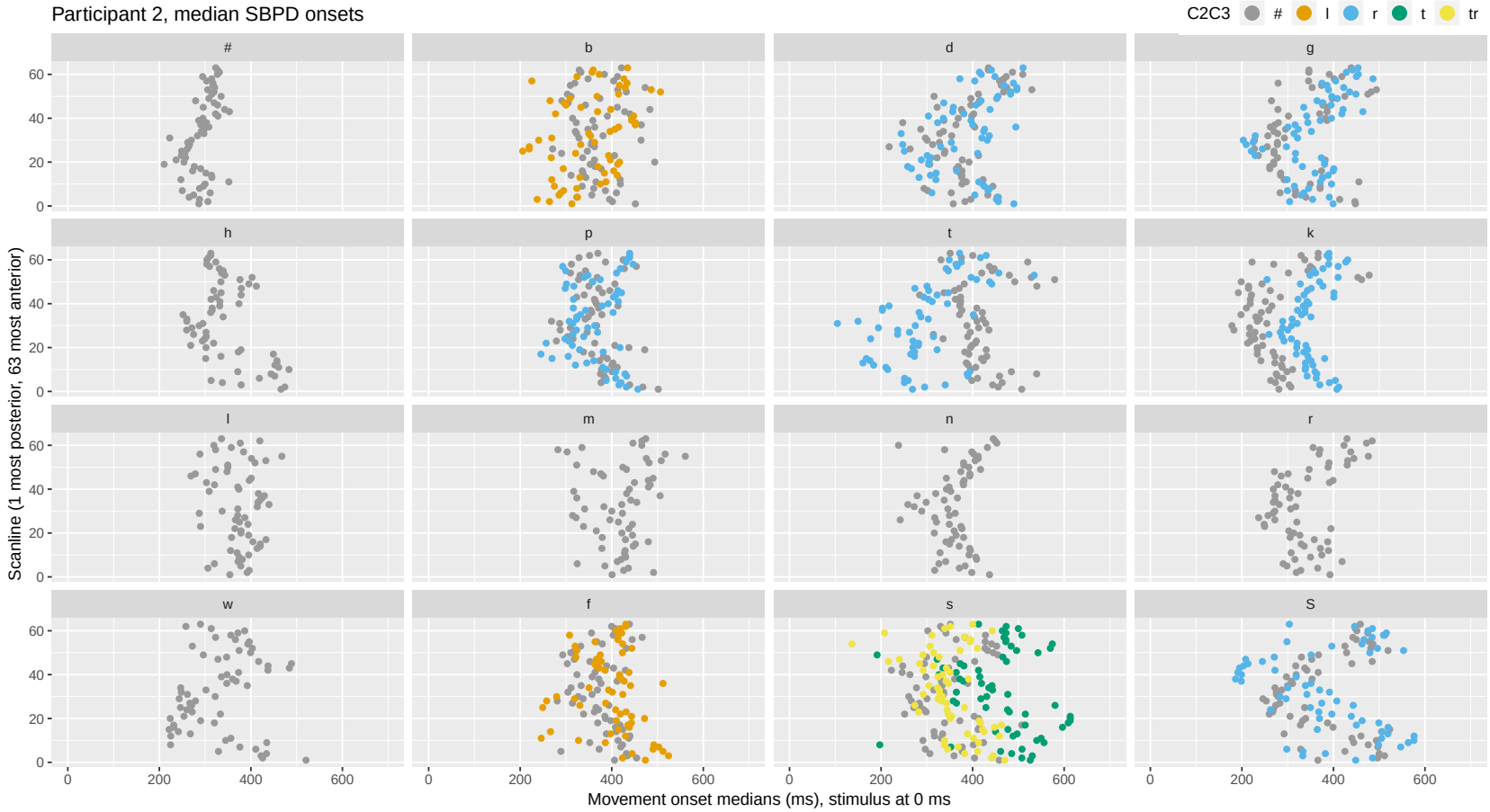

Figure 5.7: Medians of localised movement onset latencies for participant P2. Y-axis is time and x-axis position from back (1st scanline) to front (63rd scanline) conditioned by initial consonant (# marks no onset, that is, a /VC/ word). Grey dots correspond to /CVC/ utterances and coloured ones to utterances with complex onsets.

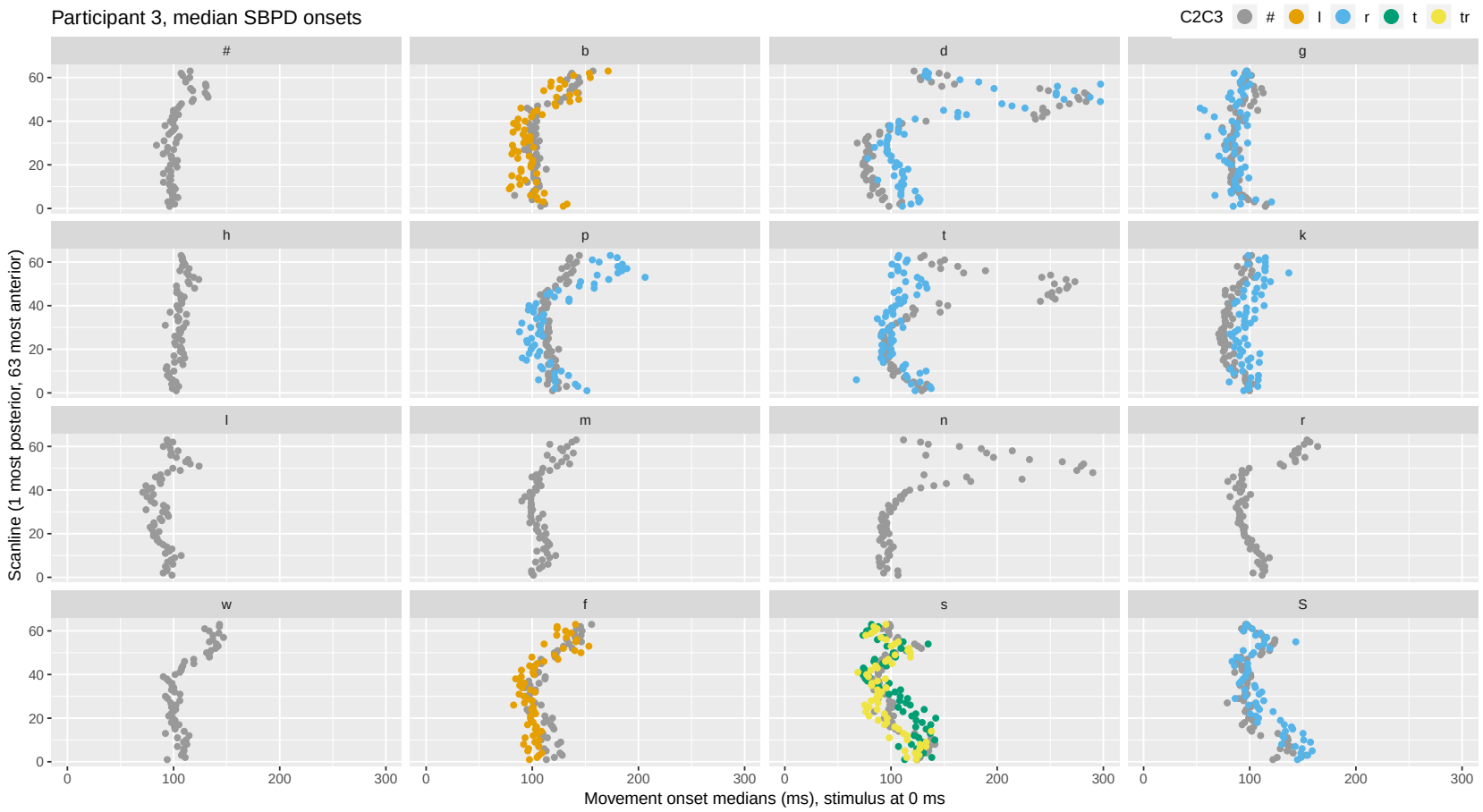

Figure 5.8: Medians of localised movement onset latencies for participant P3. Y-axis is time and x-axis position from back (1st scanline) to front (63rd scanline) conditioned by initial consonant (# marks no onset, that is, a /VC/ word). Grey dots correspond to /CVC/ utterances and coloured ones to utterances with complex onsets.

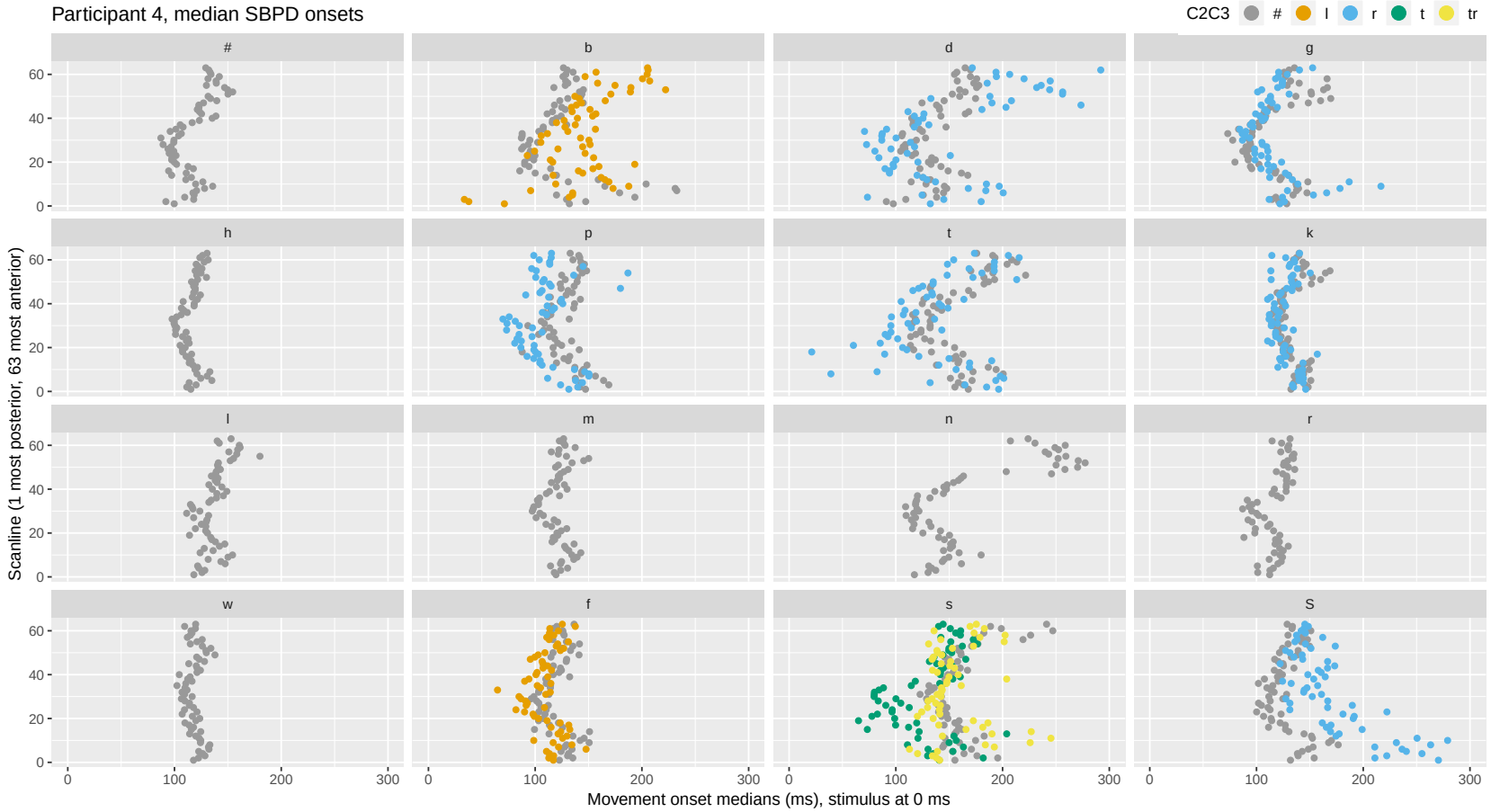

Figure 5.9: Medians of localised movement onset latencies for participant P4. Y-axis is time and x-axis position from back (1st scanline) to front (63rd scanline) conditioned by initial consonant (# marks no onset, that is, a /VC/ word). Grey dots correspond to /CVC/ utterances and coloured ones to utterances with complex onsets.

Complex onset – /CCVC/ and /CCCVC/ – words show individual interactions with the simple onset words. The interactions are difficult to analyse conclusively, highlighting the need to record more data in the future to sample the possible /CC/ and /CCC/ combinations.

5.6 Discussion

The data set of this experiment forms a relatively large UTI corpus. While analysis of the longitudinal aspects of the data remains future work, the fact that the recordings spanned between 1 and 11 months for those participants who completed the whole set, does add to the data sets potential usefulness.

The SBPD results on localised movement onsets illustrated in Figures 5.6 – 5.9, demonstrate a novel way of analysing at UTI data in a speech initiation or reaction time context. The results show that there is considerable inter- and intra-speaker variation in speech initiation strategies. While the method clearly does have room for improvement – at the moment there are a considerable number of errors in the movement onset that SBPD detects – it does already in this implementation retain a reasonable amount of the data. Operationalising articulatory reaction time in this way facilitates analysis of large UTI corpora, such as this experiment's data set without the need for time-consuming manual annotation of movement onsets.

The statistical analysis presented above forms the basis for answering the research questions about the relative timing of articulatory and acoustic onsets in speech initiation. The result on rhyme duration predicting AAI via a positive correlation pattern suggest that speech or articulation rate affects AAI: faster production of the rhyme equates to faster production of the word and according to the statistical modelling the result of this is a shorter AAI. Implications of this finding are discussed further in Chapter 7.

As expected, the results show that the inverse correlation between onset duration and acoustic reaction time actually arises from period of silent articulation preceding – that is, the AAI – preceding the acoustic onset. The results

further show that the acoustic rhyme – the VC part of the target words, including the final burst – is positively correlated with AAI.

The AAI for vowel onset words is long in the data of this experiment. This implies that a slot is left for a consonant, that is not filled. One possibility is that the articulators' movement to the location needed to generate the vowel take the same length of time regardless of whether there is an initial consonant in the utterance. However, vowel acoustics are not generated within that transitional phase, in order to avoid the impression of a glide or diphthong. This might be achieved by filling the consonant slot with a consonant whose acoustic duration is zero – that is a glottal stop.

Whatever the mechanism of filling that slot is, the results of this experiment are in disagreement with those of Mooshammer et al. (2012) who performed a similar study but with EMA instead of UTI. As was discussed in Section 2.2.2, in their acoustic results vowel onset words pattern most closely with plosive onset words. Given this if the locus of the inverse correlation pattern is in the AAI, we would expect articulatory onset times to be unaffected by the duration of the onset consonant. However, Mooshammer et al. (2012) report longer articulatory reaction times for vowel onset words than consonant onset words.

A dual modality data set with one speaker producing the same set of materials was recorded in an effort to solve this discrepancy. It was recorded on two occasions: Once in UTI and once in EMA. This experiment is the topic of the next chapter.

The long span of the recordings gives the opportunity to test whether each individual has their own habitual rest position and speech preparation pattern. However, since tongue surface contours have not yet been extracted from the data, analysing the rest positions remains future work.

Chapter 6

Experiment 3: Comparison of EMA and UTI

This experiment is a cross-methodological validation of Experiment 2. It addresses the discrepancy between the results of Experiment 2 and those of Mooshammer et al. (2012). The experiment involved a Finnish speaking participant – the author – performing the task from Experiment 2, but was recorded in separate sessions with Electromagnetic Articulography (EMA) (at University of Helsinki) and Ultrasound Tongue Imaging (UTI) (at Queen Margaret University). In addition, as the participant spoke L1 Finnish the materials were adapted for the Finnish phonotactics. The EMA recordings were obtained at Faculty of Behavioural Sciences at University of Helsinki in Finland. The principal investigator of the EMA part of this experiment was Dr. Juraj Šimko. The UTI data was recorded using the same general protocols as in Experiment 2 (Chapter 5) in the speech lab at CASL at Queen Margaret University.

6.1 Introduction

To test the assumptions in experiment 2 and in order to follow the original experiment by Rastle et al. (2005), we selected the target words to be phonotactically legal Finnish syllables of types /V/, and /CV/. In word initial syllables

the vowels can occur as short, long or as part of diphthongs. For the current experiment, we used only short monophthongs.

Finnish phonological system has eight vowels /ɑ, e, i, o, u, y, æ, ø/ and up to seventeen consonants eleven of which – namely /h, j, k, l, m, n, p, r, s, t, ʋ/ (Suomi et al. 2008, page 25) – occur in all synchronic varieties of Finnish. To provide good balance between coverage of the possible sounds and tractable length of experiment, we used these eleven core consonants and all of the vowels in our experiment.

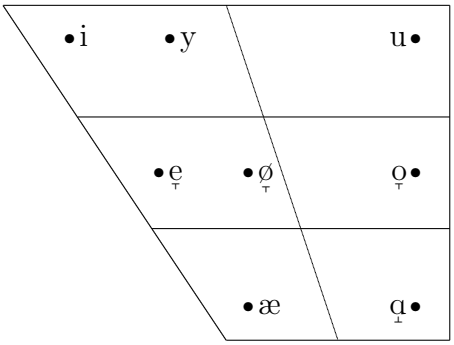

Figure 6.1: Finnish vowels on the vowel quadrilateral after Suomi et al. (2008).

6.1.1 Design of the EMA experiment

EMA is very different from UTI in terms of practical limitations of the method. Gluing the coils on the tongue and lips is a more strenuous operation in comparison with fitting the UTI helmet on. On the other hand, once the coils have been glued in place it is not too difficult to wear them for a prolonged time – a practical limit based on previous experience with the facility in Helsinki was about two hours of recording time after the initial setup. When this extended recording time is combined with the fact that the EMA system records the data practically in real time – that is, unlike with UTI there is no down time between single trials – it was possible to perform an experiment where a single session could last up to two hours. Thus, it was possible to record enough tokens to match the sample numbers in Experiment 2 in a single session. However, since

a larger data set was required when covering all of the Finnish vowels (see below) the recordings were split over two sessions. Ethical approval was obtained according to the protocols in place at University of Helsinki.

6.1.2 Set up of the EMA experiment

The EMA data was recorded at speech laboratory of the Institute of Behavioural Sciences at University of Helsinki in collaboration with Dr. Juraj Šimko and Ms Mona Lehtinen. We used the Carstens AG500 electromagnetic articulograph (Carstens Medizinelektronik 2015) in a sound-proofed recording studio.

The experimenters observed sensor movement in real-time as it was recorded. Audio was recorded simultaneously with the kinematic data using 2 microphones (AKG). One microphone was attached to the EMA control laptop (DELL Latitude D830); the other was used to record higher quality audio on a Mac computer with ProTools recording software (version 10). Auditory prompts for the participant were presented through headphones (Sennheiser, HD250, Linear II). The experimental setup (prompts) ran on a PC (HP Compaq 8200 Elite) and was created and run with Presentation®-software (Neurobehavioral systems 2015).

Three EMA sensor coils were used for correction of head movements and five were used for recording articulatory data. The head correction coils were placed on relatively immobile points on the participants skin: one behind each ear and one on the ridge of his nose. Articulatory data was recorded from coils on the upper and lower lip, the lower incisors to capture jaw movement and three coils on the tongue: one as far back on the midline as was comfortably possible (referred to as 'tongue body' below), one as near the tongue tip on the midline as was comfortably possible ('tongue blade'), and the third one about half-way between the two other coils ('tongue dorsum'). The trajectory signals, recorded at a sampling rate of 200 Hz, were processed using Tapad system (Hoole and Zierdt 2010) and subsequently head-movement corrected, smoothed by an 8- point Bartlett window and up-sampled to 1000 Hz using cubic spline interpolation.

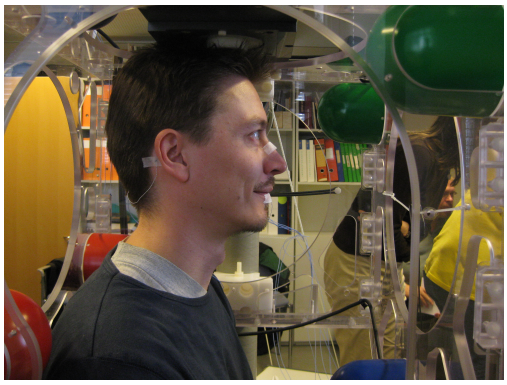

Figure 6.2: The author in the AG500. (The picture was taken before the system was installed in the sound-proofed studio.)

6.1.3 EMA data quality and postprocessing

The raw data generated by EMA is very noisy and needs a to be post-processed to provide analysable data. The post-processing was done by Dr. Šimko, who also performed the semi-manual movement onset detection with a Matlab script that he implemented himself.

The EMA data were processed in a standard way, that is, from the raw recordings we calculated x, y and z coordinates for lip (upper and lower), tongue blade, tongue dorsum, tongue body, and jaw, in time. The data were head corrected using the three sensors placed on the nose and behind ears. Regions in the EMA recordings were selected for further analysis based on the acoustic annotations made by the author. This involved aligning the low grade audio recording made by the EMA machine and a separate hi-fi audio recording This was done by aligning the audio markers recorded with a direct cable from the computer running the Presentation software. Crucially, all synchronisation signals between different parts of the laboratory setup were recorded by directly connected cables to avoid any offsets in synchrony.

For purposes of identifying movement onsets, four channels were considered. Three of these were the two recording coils for tongue and the jaw. The fourth channel was the Euclidean distance between the upper and lower lip recording coils. The onset of the movement on each channel was identified

from the last three local tangential velocity minima before maximum tangential velocity for the onset gesture, which was identified by relating the movement data to the acoustic annotations. The script offers three minima candidates – displayed on a representation of the movement trajectory of the channel being analysed, and prompts the human annotator to choose the most appropriate of them.

6.2 Materials and methods

6.2.1 Participant

The participant was a native male speaker of Finnish – the author – with corrected to normal vision and no known hearing or speech problems participated in the experiment (age 38 years).

6.2.2 Stimuli

We carried out a partial replication of the Rastle et al. (2005) delayed naming experiment with the following changes: Instead of using phonetically transcribed syllables as stimuli, we used orthographically transcribed Finnish syllables. In the UTI experiment orthographic 'ä' was substituted with 'ae' and 'ö' with 'oe' to avoid character encoding problems while running the experiment.

The purpose of this experiment was to see if words with a vowel onset pattern in a systematic way with those with a consonant onset and if there is a difference in how EMA and UTI register the articulatory onset in vowel onset tokens. Thus, the syllables were of /CV/ and /V/ type. The syllables were chosen to span the phonotactically permissible initial syllables of the given types for Finnish. The vowel list consisted of the complete list of Finnish short vowels: [ɑ], [e], [i], [o], [u], [y], [æ] and [ø], while the consonant list was restricted by the Finnish phonotactics to be [h], [j], [k], [l], [m], [n], [p], [r], [s], [t] and [v]. These sounds can be combined to produce 88 syllables of the /CV/ type and 8 syllables of the /V/ type giving a total of 96 target syllables.

6.2.3 EMA procedure

Each target syllable was shown to the participant on a computer screen. After two seconds, the visual go-signal – angle brackets – appeared around the target: for example, 'la' changing into '>la<'. The participant was instructed to read the stimulus out loud, as soon and as accurately as possible. Being the sole participant, the author knew that it was important remain at rest and not make any articulatory movements until he saw the go-signal. The 96 stimuli were repeated five times throughout the whole experiment in five internally randomised blocks. The randomisation was constricted to guarantee that there were no repeats at block boundaries (which would be caused by one block ending with the token the next one begins with).

6.2.4 UTI procedure

The UTI data was recorded in five batches of about 20 minutes. The batches were spread over two days, which were about a year and a half apart. At the time of writing analysis results of only the first day of UTI recording – first two batches are available. The results below reflect this in that the UTI data set does not cover all of the onsets and vowel qualities available in the EMA data.

Each target syllable was shown to the participant on a computer screen. After the participant indicated with a keypress that he was ready to begin the trial, the system played the go-signal (a 1 kHz beep) delayed by a random interval between 1.2 and 1.8 seconds. The participant was instructed to read the stimulus out loud, as soon and as accurately as possible. Again, being the sole participant, the author knew the importance of remaining at rest before he observed the go-signal. The 96 stimuli were repeated five times throughout the whole experiment in five internally randomised blocks. The randomisation was automatically checked to guarantee that there were no repeats at block boundaries.

6.3 Audio analysis

Following the same procedure as in Experiment 2, FAVE (Rosenfelder et al. 2011) was used to produce a first approximation of phone segmentation. The raw segmentation produced by FAVE was manually corrected in Praat (Boersma and Weenink 2010). The spectrogram was calculated with a 5 ms window on a range of 0-7000 Hz with auto-scaling turned off (standard setting is to have it on).

6.3.1 Audio segmentation rules

The segmentation rules below are an adaptation of those set out in Experiment 2 based on both the productions of the participant and descriptions given by Suomi et al. (2008).

All onsets were frequently preceded by transient smacking noises produced by the participant opening his mouth and lowering his tongue from the palate. These were defined as not being part of the acoustic speech even though they seemed to be more frequent with some sounds (such as /l/) than others (for example, the vowels).

Vowels: In /V/ tokens, vowel onset was defined either as the plosive like burst (likely an indication of glottal release) preceding start of phonation or the start of phonation as determined from the waveform. In /CV/ tokens, the definition of vowel onset was specific to the type of consonant and rules are listed below. Utterance end was not extracted and therefore vowel offsets were not marked.

/h, s/**:** Onset was defined as the onset of frication noise in the spectrogram. This usually occurred around 2 kHz for **/h/** and around 2.5 kHz for **/s/**, or as a burst across the whole spectrum. Consonant offset and vowel onset was defined as the onset of phonation as observed on the waveform.

/j, l, ʊ/**:** Onset was defined as start of phonation and offset either as a dip in waveform amplitude and/or shift into a more vowel like formant structure in the waveform

/m, n/**:** Onset was defined as the onset of phonation being visible on the waveform. Consonant offset / vowel onset was primarily determined by a dip in amplitude of the waveform and secondarily as a change in the quality of sound visible both as a change in the form of the pulses in the waveform and a shift of formants in the spectrogram. Occasionally a weak but clear release burst was evident on the spectrogram and/of waveform and this was used to pinpoint the demarcation point.

/k, p, t/**:** Finnish voiceless plosives are unaspirated and unvoiced. So, onset was defined as the onset of the release burst and offset / vowel onset as the beginning of phonation both determined on the waveform.

/r/**:** Onset was defined either as the onset of phonation or onset of significant, continuous frication. Offset was defined as either a dip in waveform amplitude or more commonly as the point when formants started to shift from their original positions towards the vowel targets.

6.4 Results

Before thresholding, the UTI set consisted of 479 recorded tokens. As we see in Table 6.1, the syllable 'ja' was repeated 9 times instead of the standard 5, and syllables 'na', 'hu', 'ju', 'pu', and 'tu' each missed one trial each.

Table 6.1: Number of tokens recorded in UTI listed by onset consonant (C) and nucleus vowel (V). # denotes no onset consonant referring to /V/ syllables.

V \ C	/# /	/h/	/j/	/k/	/l/	/m/	/n/	/p/	/r/	/s/	/t/	/v/
/ɑ/	5	5	9	5	5	5	4	5	5	5	5	5
/æ/	5	5	5	5	5	5	5	5	5	5	5	5
/e/	5	5	5	5	5	5	5	5	5	5	5	5
/i/	5	4	5	5	5	5	5	5	5	5	5	5
/o/	5	5	5	5	5	5	5	5	5	5	5	5
/ø/	5	5	5	5	5	5	5	5	5	5	5	5
/u/	5	4	4	5	5	5	5	4	5	5	4	5
/y/	5	5	5	5	5	5	5	5	5	5	5	5

Also before thresholding, the EMA data set consisted of 402 recorded to-

kens. Table 6.2 lists the number of recorded tokens by onset consonant and nucleus vowel. Due to sensor tracking errors there were considerably more missing trials in EMA than UTI.

Table 6.2: Number of tokens recorded in EMA listed by onset consonant (C) and nucleus vowel (V). # denotes no onset consonant referring to /V/ syllables.

V \ C	/# /	/h/	/j/	/k/	/l/	/m/	/n/	/p/	/r/	/s/	/t/	/v/
/ɑ/	3	5	4	5	2	5	3	4	5	5	4	3
/æ/	5	5	5	5	4	3	2	5	5	4	3	3
/e/	5	5	5	4	5	5	1	5	4	5	4	5
/i/	4	4	4	3	5	5	4	5	3	5	2	4
/o/	4	3	5	5	5	3	5	4	5	4	2	5
/ø/	4	4	5	4	5	5	4	5	5	4	5	4
/u/	4	5	5	4	5	4	4	3	5	5	2	2
/y/	5	4	5	3	5	4	4	4	5	4	4	5

6.4.1 Token exclusion criteria

Tokens with early starts were removed from the data set by thresholding with the lower bound values used in Experiment 2: 28 ms for the Scanline Based Pixel Difference (SBPD) onset (as set out in Section 3.3.2 based on Chiu and Gick 2014), 58 ms for the acoustic reaction time between, and 0 ms for the Articulatory to Acoustic onset Interval (AAI). The upper bound for acoustic reaction time was dropped because the participant had fairly long reaction times. No EMA tokens were removed by thresholding and only two UTI tokens were removed by it. The removed UTI tokens were a repetition of 'su' and a repetition of 'hi'. Total number of tokens included in further analysis was 477 UTI tokens and 402 EMA tokens.

6.4.2 Comparison of reaction time measures

Linear mixed effects models were fitted to the data to predict articulatory reaction time, acoustic reaction time, and the AAI. The results mirror those of Experiment 2. In addition, a level difference between the two modalities – EMA

and UTI – was identified as part of all three models. However, syllable type – /V/ vs. /CV/ – was not found to influence the results.

The data was analysed in R (R Core Team 2013) by iteratively fitting linear models (step-up process Baayen 2008) with version 1.1-21.9000 of the `lme4` package (Bates et al. 2015) to explain the variation in acoustic reaction time, and AAI. For articulatory reaction time the procedure was otherwise the same but the models were ordinary linear models fitted with the built-in `lm` function of R because including a random effect for word lead to a singular (non-converging) model. Step-up comparisons were performed with the built-in `anova` function of R in all cases.

Articulatory reaction time

The models for articulatory reaction time were ordinary linear regression models (see explanation in the previous paragraph). The first model included only an effect for trial number and was (in R formula notation):

$$ArtRT \sim trial + (1|word). \tag{6.1}$$

Testing significance with ANOVA yields a statistically significant result: P-value $\approx 4.283 \times 10^{-9}$ evaluated on the F distribution with degrees of freedom 1 and 877. Including modality in Model 6.1 we have:

$$ArtRT \sim trial + modality. \tag{6.2}$$

Testing with ANOVA whether the second model has improved the fit yields a statistically significant result: P-value $\approx 5.552 \times 10^{-13}$ evaluated on the F distribution with degrees of freedom 1 and 876. Further models for predicting articulatory reaction time failed to reach significance.

Summary of the final model – fitted after removing outliers by culling data points with an absolute value of the normalised residual in excess of 2.5 – is given in Table 6.3. In the table we see that the model coefficients are trial: -0.15 and modality for EMA: 32.6. This means that the participant got on average

Table 6.3: Summary of the final linear model of articulatory RT.

```
Call:
lm(formula = art_RT ~ trial + modality, data =
    subset(E3, abs(scale(resid(E3.lm_art3))) < 2.5))

Residuals:
     Min       1Q   Median       3Q      Max
-153.357  -50.929   -8.623   45.717  197.371

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 250.78102    4.15316  60.383  < 2e-16 ***
trial        -0.13150    0.01715  -7.668 4.71e-14 ***
modalityEMA  32.60900    4.78110   6.820 1.71e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 65.62 on 859 degrees of freedom
Multiple R-squared:  0.08384,Adjusted R-squared:  0.08171
F-statistic: 39.31 on 2 and 859 DF,  p-value: < 2.2e-16
```

slightly faster over the course of a batch of recordings, and that the articulatory reaction times recorded in EMA were on average 32.6 ms longer than those recorded with UTI.

Acoustic reaction time

The iteration process identified a statistically significant model that successfully predicts acoustic reaction time ($AcRT$ in the formulas). The first model included only a random effect for word and was (in R formula notation):

$$AcRT \sim (1|word) \tag{6.3}$$

where $(1|word)$ is the random effect for the word. Including the trial number in Model 6.3 we have:

$$AcRT \sim trial + (1|word). \tag{6.4}$$

Testing with ANOVA whether the second model has improved the fit yields a statistically significant result: P-value < 2.2×10^{-16} evaluated on the χ^2 distribution with one degree of freedom. Including modality in Model 6.4 we have:

$$AcRT \sim trial + modality + (1|word). \tag{6.5}$$

Testing with ANOVA whether the third model has improved the fit yields a statistically significant result: P-value < 2.2×10^{-16} evaluated on the χ^2 distribution with one degree of freedom. Including Onset consonant's acoustic Duration (OD) in Model 6.5 we have:

$$AcRT \sim trial + modality + OD + (1|word). \tag{6.6}$$

Testing whether the third model has improved the fit yields a marginally statistically significant result: P-value $\approx 5.945 \times 10^{-6}$ evaluated on the χ^2 distribution with one degree of freedom.

Summary of the final model – fitted after removing outliers by culling data points with an absolute value of the normalised residual in excess of 2.5 – is given in Table 6.4. In the table we see that coefficients of the fixed effects are trial: -0.30, modality for EMA: 59.4, and OD: -0.38. It is worth noting that for this participant the trial effect has the same sign when compared with the trial effect on articulatory reaction time, meaning that for this participant both acoustic and articulatory reaction times got relatively shorter towards the end of a recording session. We also see that the average level difference between EMA and UTI is about twice that of articulatory reaction time.

Articulatory to Acoustic onset Interval

Similarly to acoustic reaction time, the iteration process identified a statistically significant model that successfully predicts AAI. The first model was:

$$AAI \sim (1|word) \tag{6.7}$$

Table 6.4: Summary of the final acoustic reaction time mixed effects model.

```
Linear mixed model fit by REML ['lmerMod']
Formula: ac_RT ~ trial + modality + OD + (1 | word)
   Data: subset(E3, abs(scale(resid(E3.lmer_ac4))) < 2.5)

REML criterion at convergence: 9731.6

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.39523 -0.72388 -0.05376  0.63443  2.82441

Random effects:
 Groups   Name        Variance Std.Dev.
 word     (Intercept)  196.2   14.01
 Residual             4652.8   68.21
Number of obs: 860, groups:  word, 120

Fixed effects:
             Estimate Std. Error t value
(Intercept) 494.20984    6.46754  76.414
trial        -0.29602    0.01811 -16.345
modalityEMA  59.42424    5.10851  11.632
OD           -0.37875    0.07452  -5.082

Correlation of Fixed Effects:
            (Intr) trial  mdlEMA
trial       -0.520
modalityEMA -0.151 -0.349
OD          -0.711  0.087 -0.002
```

where $(1|word)$ is the random effect for the word. Including the trial number in Model 6.7 we have:

$$AAI \sim trial + (1|word). \tag{6.8}$$

Testing whether the second model has improved the fit yields a statistically significant result: P-value $< 2.2 \times 10^{-16}$ evaluated on the χ^2 distribution with one degree of freedom. Including modality in Model 6.8 we have:

$$AAI \sim trial + modality + (1|word). \tag{6.9}$$

Testing whether the second model has improved the fit yields a statistically significant result: P-value $\approx 1.328 \times 10^{-11}$ evaluated on the χ^2 distribution with one degree of freedom. Including OD in Model 6.9 we have:

$$AAI \sim trial + modality + OD + (1|word). \tag{6.10}$$

Testing whether the second model has improved the fit yields a statistically significant result: P-value $\approx 1.297 \times 10^{-14}$ evaluated on the χ^2 distribution with one degree of freedom.

Summary of the final model – fitted again after removing outliers by culling data points with an absolute value of the normalised residual in excess of 2.5 – is given in Table 6.5. In the table we see that the coefficients of the fixed effects are trial: -0.15, modality for EMA: 25.7, and OD: -0.45.

We see that these results are in line with those of articulatory and acoustic reaction times. AAI has a trial effect of the same size as the same effect on articulatory reaction time and of the same sign. This means that the trial effect on acoustic reaction time is the result of the effect on AAI being added together with the effect on articulatory reaction time. The same seems to hold true for the level difference between EMA and UTI as the difference for AAI is of roughly the same size as that for articulatory reaction time or half of the effect for acoustic reaction time.

Table 6.5: Summary of the final Articulatory to Acoustic onset Interval mixed effects model.

```
Linear mixed model fit by REML ['lmerMod']
Formula: AAI ~ trial + modality + OD + (1 | word)
   Data: subset(E3, abs(scale(resid(E3.lmer_aai4))) < 2.5)

REML criterion at convergence: 9085.7

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.49397 -0.69657 -0.02879  0.64748  2.67672

Random effects:
 Groups   Name        Variance Std.Dev.
 word     (Intercept)  114.6   10.70
 Residual             2063.2   45.42
Number of obs: 864, groups:  word, 120

Fixed effects:
             Estimate Std. Error t value
(Intercept) 243.33763    4.38661  55.473
trial        -0.15207    0.01201 -12.665
modalityEMA  25.70436    3.38926   7.584
OD           -0.45671    0.05069  -9.010

Correlation of Fixed Effects:
            (Intr) trial  mdlEMA
trial       -0.510
modalityEMA -0.170 -0.333
OD          -0.711  0.085  0.008
```

Figures 6.3–6.5 show data from both EMA and UTI side by side. As can be seen (Figure 6.3) acoustic reaction times are essentially the same despite the fact that the go-signal was auditory in UTI and visual in EMA. Both repeat the inverse correlation patterns of acoustic reaction time and AAI with OD already seen in Experiment 2, while there is no such pattern present in either data set for articulatory reaction time and OD. The only exception is that in the EMA data the vowel onset syllables do not seem to fit the pattern (Figure 6.5). However, the above statistical analysis does not support this informal observation.

6.4.3 Location of movement onset

Location of movement onset was studied to check if the level difference between EMA and UTI could be a result of the (by necessity) fairly frontal placement of the EMA sensors in comparison with the field of view of UTI. Figure 6.6 shows the onset locations in the UTI data. The vowel onset tokens and most of the consonant onset tokens show early activation at the back of the tongue in regions that did not have EMA sensors on them. However, the posterior onset pattern is not totally clear. Especially in the case of /s/, the data seems unclear with many anterior scanlines registering an earlier onset than the posterior half of scanlines.

Table 6.6 lists the frequencies for each EMA sensor being the first to register movement and Table 6.7 lists the relative proportions of each onset location for each onset consonant and the data as a whole. We see that 37.4 % of the onsets are registered by the tongue dorsum sensor, with the tongue body sensor which is further back on the tongue registering 22.4 % of onsets as well. Lip aperture registered as the onset location in 20.6 % of the total data.

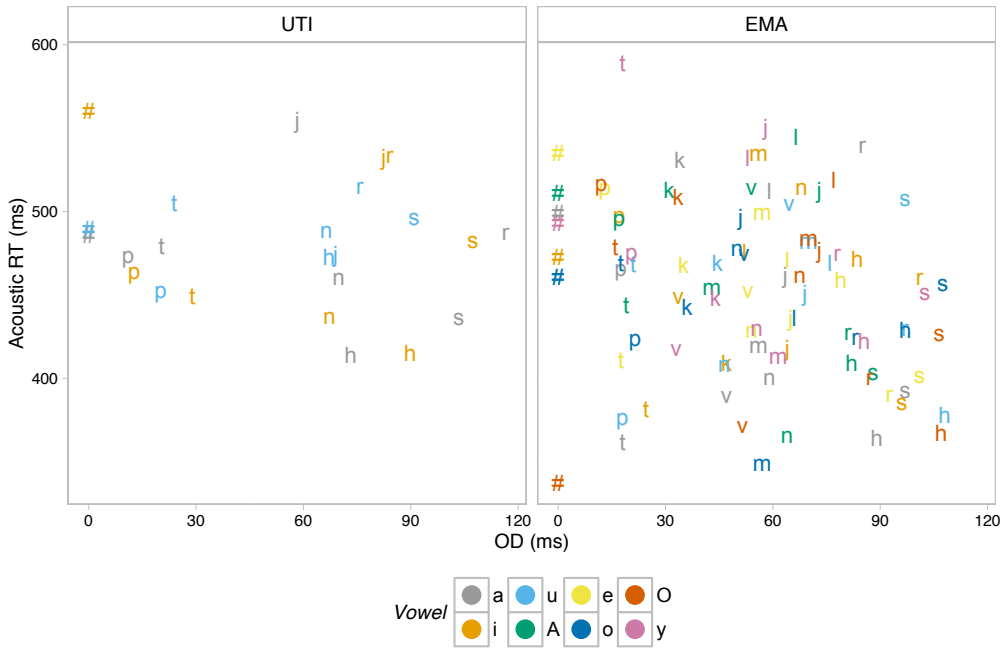

Figure 6.3: Acoustic reaction times as a function of OD from EMA and UTI data.

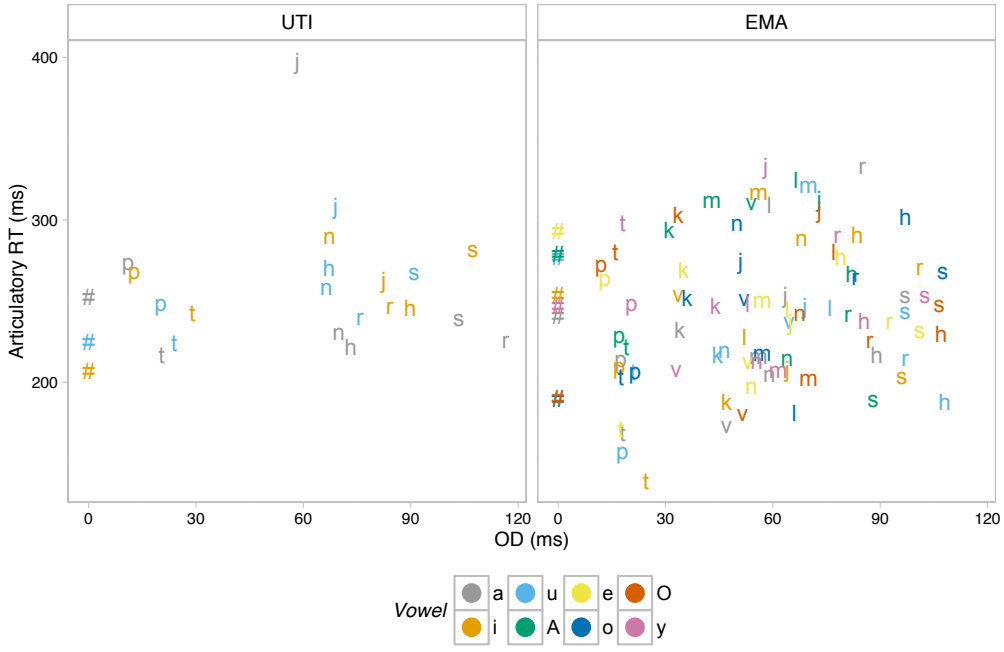

Figure 6.4: Articulatory reaction times as a function of OD from EMA and UTI data. EMA reaction times are based on all sensors.

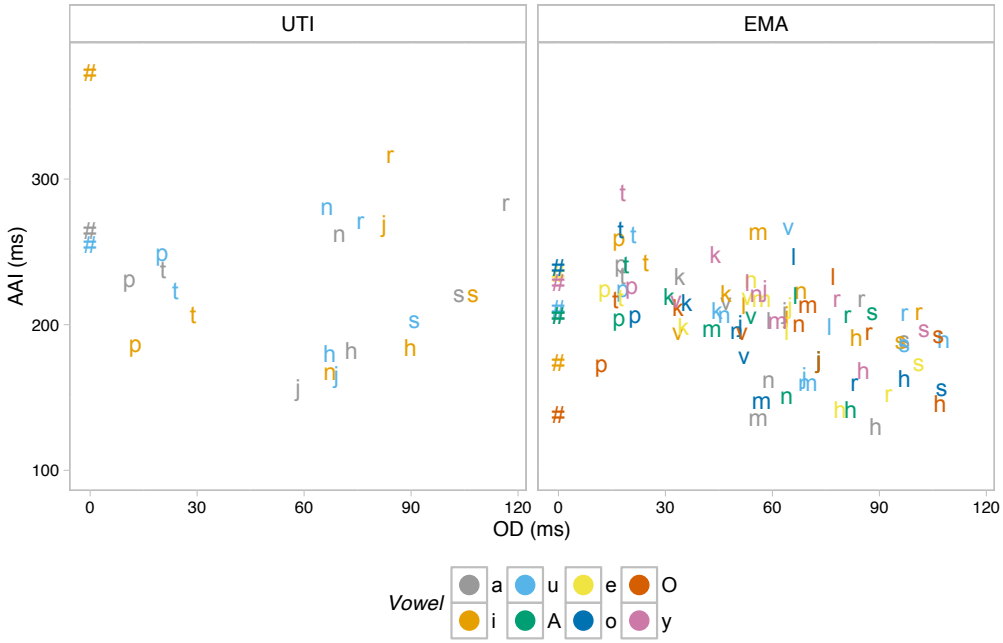

Figure 6.5: AAIs as a function of OD from EMA and UTI data. EMA AAIs are based on all sensors.

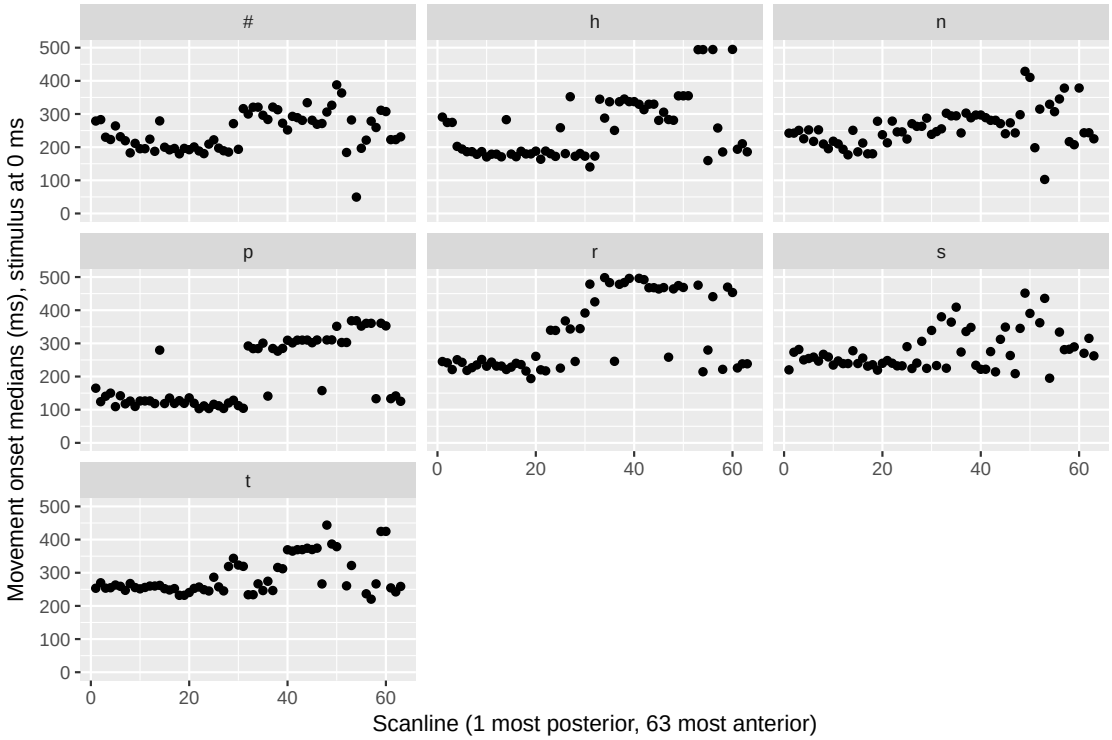

Figure 6.6: Medians of localised movement onset latencies from SBPDs.

Table 6.6: Onset location versus onset consonant in EMA: Frequencies of tokens where a given EMA sensor is the first to register movement cross-tabulated against onset consonant.

	/#/	/h/	/j/	/k/	/l/	/m/	/n/	/p/	/r/	/s/	/t/	/v/	Sum
jaw	2	4	6	1	3	1	4	1	5	4	2	5	38
lip aperture	7	9	7	4	7	1	1	1	**11**	**18**	2	**15**	83
tongue blade	4	1	2	3	7	1	**7**	3	5	2	5	1	41
tongue dorsum	**20**	**12**	**17**	**16**	**9**	**16**	**8**	**19**	9	9	7	8	150
tongue body	1	9	6	9	**10**	**15**	**7**	11	7	3	**10**	2	90
Sum	34	35	38	33	36	34	27	35	37	36	26	31	402

Table 6.7: Onset location versus onset consonant in EMA: Proportions of tokens where a given EMA sensor is the first to register movement cross-tabulated against onset consonant. Numbers given are percentages of the total number of analysed tokens with each onset. The last column gives the total percentages over the whole EMA data.

	/#/	/h/	/j/	/k/	/l/	/m/	/n/	/p/	/r/	/s/	/t/	/v/	Total
jaw	5.9	11.4	15.8	3.0	8.3	2.9	14.8	2.9	14.0	11.1	7.7	16.1	9.5
lip aperture	20.6	25.7	18.4	12.1	19.4	2.9	3.7	2.9	**30.0**	**50.0**	7.7	**48.4**	20.6
tongue blade	11.8	2.9	5.3	9.1	19.4	2.9	25.9	8.6	14.0	5.6	19.2	3.2	10.2
tongue dorsum	**58.8**	**34.3**	**44.7**	**48.5**	**25.0**	**47.1**	**29.6**	**54.3**	24.0	25.0	26.9	25.8	**37.3**
tongue body	2.9	25.7	15.8	27.3	**27.8**	**44.1**	25.9	31.4	19.0	8.3	**38.5**	6.5	22.4

6.5 Discussion

The results of this experiment fail to show that the discrepancy between the results of Experiment 2 and Mooshammer et al. (2012) would be a result of different articulatory measurement methods. Thus, the question of this discrepancy remains open and will be discussed further in the next chapter.

The results of this chapter do, however, show that EMA registers longer reaction times in both articulation and acoustics. Part of the effect on acoustic reaction time can be attributed to the AAI, which could in part be a result of slower articulation rate, but without corresponding articulation rate data this can not be checked.

Regardless of articulation rate, part of the effect is a later registration of articulatory reaction time. A possible explanation for this would have been movement onset location. However, since analysis of location of articulatory onset in UTI proved inconclusive, we can only conclude that the situation merits further study. After all, there are other possible explanations for the discrepancy between EMA and UTI results. To mention two, one possible explanation is the difference in modalities of the go-signals: EMA used a visual signal and UTI an auditory one. Another is that the measurement apparatus for each recording method may have affected the participant's speech production in different ways – one impeding response speed more than the other. Properly answering this question clearly requires more data to be recorded. This remains future work.

In the next chapter we will discuss the research questions in light of the results of the experiments presented, followed by a discussion of the new analysis methods and a more general discussion.

Chapter 7

Discussion

As stated earlier, this thesis has two main goals. First, to analyse pre-speech tongue movements and relate their timing to the timing of the whole utterance. Second, to facilitate reaching the first goal and as a contribution to future research, new quantitative, efficient and replicable analysis methods of Ultrasound Tongue Imaging (UTI) data have been developed.

The next two sections show how the thesis has met its goals and then proceed with a more general discussion of the results. The next section revisits the research questions in light of the results of the experiments and discusses some immediate implications of the results. It is followed by a section evaluating the methods developed in this thesis.

7.1 Research questions answered

7.1.1 Research Question 1: Timing of utterance onsets

Since Question 1 is a general question that is answered by answering the more specific questions 1a, 1b, and 1c, we will first answer each of these in turn and then return to the main question.

Question 1a: Is articulatory reaction time affected by the acoustic duration of the onset consonant (OD) or by the acoustic duration of the utterance's rhyme?

It was hypothesised that articulatory reaction time measured from the tongue is expected to depend on Onset consonant's acoustic Duration (OD) but not on the duration of the utterance's rhyme. As we have seen in Section 5.5.1, the statistical models agree with the second part original hypothesis, but not the first: articulatory reaction time is not affected by OD nor by rhyme duration.

However, just because statistical models fail to reach significance, does not mean that articulatory reaction time does not have structure affected by the phonetic content of the target utterance, but it does mean that such structure – if it even exists – is dwarfed by the size of the effect of the utterances phonetic timing on Articulatory to Acoustic onset Interval (AAI). We conclude that according to the data of this thesis, in delayed naming with the Rastle's instructions, articulatory reaction time is a noisy constant that depends on neither the OD nor the acoustic rhyme duration.

Question 1b: Is acoustic reaction time affected by the acoustic duration of the onset consonant (OD) or by the acoustic duration of the utterance's rhyme?

It was hypothesised that acoustic reaction time was would be inversely correlated with OD and positively correlated with rhyme duration. The statistical analysis in Section 5.5.1 confirms both parts of this hypothesis.

More specifically, acoustic reaction time has a strong inverse linear dependence to OD and a positive, but smaller in magnitude, linear dependence to acoustic rhyme duration. The first part of the result is in line with the data reported by Rastle et al. (2005) and the second part is a novel finding.

Question 1c: Is the Articulatory to Acoustic onset Interval (AAI) affected by the acoustic duration of the onset consonant (OD) or by the acoustic duration of the utterance's rhyme?

It was hypothesised that the AAI is expected to correlate inversely with OD and positively with rhyme duration. This hypothesis was also confirmed by the statistical analysis in Section 5.5.1.

Like the acoustic reaction time, the AAI has a strong inverse linear dependence OD and a positive, but slightly smaller in magnitude, linear dependence to acoustic rhyme duration.

Question 1: What is the relative timing (and absolute reaction time in relation to the go-signal) of tongue movement initiation (or articulatory reaction time) and acoustic initiation (or acoustic reaction time) in different phonetic contexts in a speech reaction time task, following instructions used by Rastle et al. (2005)?

First, articulatory reaction time is a noisy constant. Second, the inverse correlation of AAI and OD explains the inverse correlation of acoustic reaction time and OD. Third, the AAI has a positive linear dependence on the acoustic duration of the rhyme, meaning that the AAI should be considered part of the word.

The answers of the research questions have lead to the formulation of a conceptual model of utterance initiation in delayed naming with Rastle's instructions. The model is represented in Figure 7.1. We see that the articulatory reaction time ends when the participant initiates movement from rest position. This is directly followed by speech articulation. The speech articulation is broken down into consonant closing and releasing gestures, followed by vowel articulation marked as 'vowel trough' because, the acoustic vowel typically corresponds to a period with lower values of Pixel Difference (PD). Importantly, the acoustic consonant reflects the articulatory gestures at a lag that depends on the duration of the consonant (which in turn depends on the phonetic identity of the consonant) – acoustically longer consonants have an earlier acoustic onset than acoustically shorter consonants. This makes acoustic reaction time a complex measure that is affected by articulatory reaction time, speech rate, and the acoustic duration of the onset consonant.

The way that the consonants behave acoustically according to this model is very similar to how consonants behave articulatorily according to the C-center theory in Articulatory phonology as discussed in Section 2.1.6 and by Browman and Goldstein (1988). While there is a clear similarity, it should be kept in

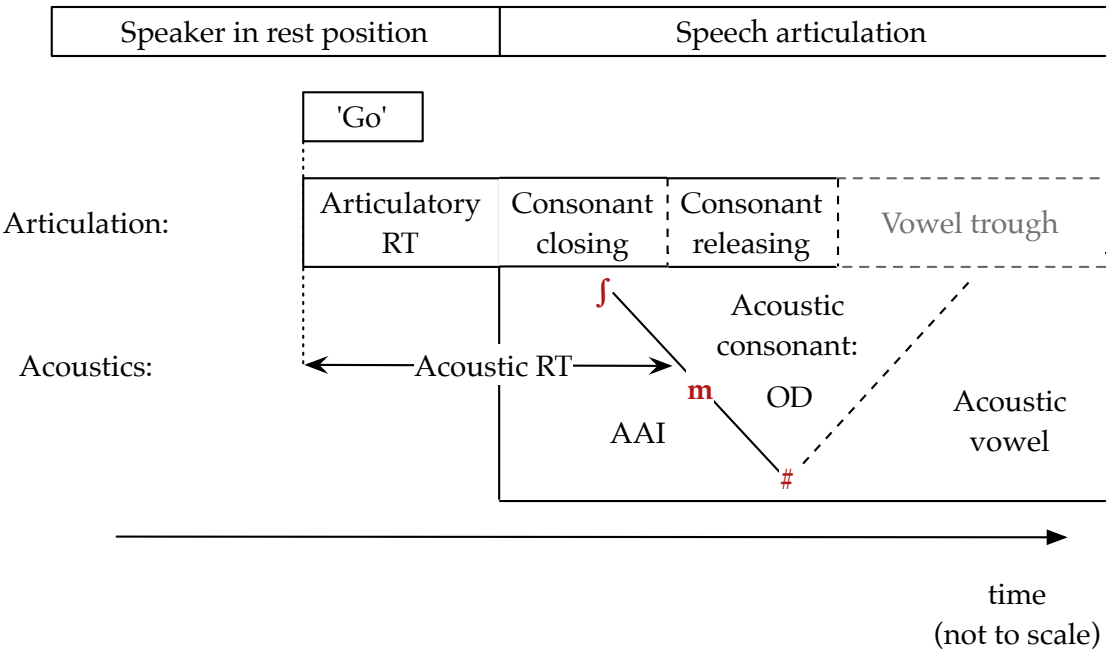

Figure 7.1: Delayed naming timeline with an account for the inverse correlation results. In this model acoustic reaction time is a complex measure which is affected by acoustic duration of the first segment(s), and by articulation rate, because the AAI period is part of the articulatory word. Dashed lines indicate boundaries that were not specifically studied in this thesis.

mind that it is not directly evident how the articulatory prediction of C-center relates to the acoustic relationship in the model in Figure 7.1. This matter clearly provides an interesting topic for future research.

For the /VC/ words, the data from Experiments 2 and 3 clearly shows – and the model in Figure 7.1 also reflects this – that they pattern as if they had an onset consonant whose acoustic duration is 0 ms. However, it remains unclear why this is so. A possible explanation is that the speakers produce an unaspirated glottal plosive – effectively a consonant whose acoustic duration is 0 ms – before the onset of the acoustic vowel, but this can not be confirmed nor disproved with the current results.

Another possible explanation is that the onset consonant production is superimposed on the slow vowel production, but that the articulation of both is initiated at the same time as proposed by the coproduction model of coarticula-

tion (Farnetani and Recasens 1999). If this were the case, then the long acoustic latency of vowels would be explained by the time it takes for the slow vowel gesture to reach a position where the oral configuration is close enough to the correct configuration for the vowel, that sound output can be initiated without compromising intelligibility. This hypothesis is also a possible explanation for the discrepancy between vowel onset words in Electromagnetic Articulography (EMA) and UTI, which is the topic of the next section.

7.1.2 Research Question 2: Difference between EMA and UTI

Do Ultrasound Tongue Imaging (UTI) and Electromagnetic Articulography (EMA) provide the same view of articulatory onset independent of the phonetic onset of an utterance?

It was hypothesised that there is a difference in how the two measurement methods see articulatory reaction time of vowel onset words. However, this hypothesis is not supported by the statistical analysis in Section 6.4.2. The analysis shows a difference between articulatory onset time derived from Ultrasound Tongue Imaging (UTI) and Electromagnetic Articulography (EMA), but the effect is global bias and not specific to vowel onset words nor to articulation: In the data of Experiment 3, EMA shows on average longer articulatory and acoustic reaction times than UTI, but utterance type – /V/ vs. /CV/ does not affect this phenomenon.

Based on the results of Experiment 3, we can thus conclude that the apparent discrepancy between data from Experiment 2 and the results of Mooshammer et al. (2012) is attributable to other differences between the experiments. In particular, without experimental data directly comparing the articulatory consequences of the Rastle task and the task used by Mooshammer et al. (2012) in combination with the syllable structure variation the latter used, we can not rule out the possibility of there being further confounding effects from the task and from syllable structure.

If we consider Figures 5.6 - 5.9 and Figure 6.6, we see that for both /VC/ and /CVC/ words the Scanline Based Pixel Difference (SBPD) median onset

curves all show an early activation in the scanlines corresponding to the back of the tongue. While the /V/ tokens do show earlier activation in Figure 6.6, the differences are a continuum, not a categorical difference between /V/ and /CV/ tokens.

A possible explanation for this is the consonant/vowel coproduction model (Farnetani and Recasens 1999) which fits the vowel onset data as discussed at the end of the previous section. This would also make physiological sense, because to articulate a vowel, a speaker needs to move the body mass of the tongue and that takes time. So, regardless of whether the speaker is initiating a vowel or consonant onset utterance, they need to activate substantial number of muscle fibres producing the kind of strong activation as seen in PD at the beginning of an utterance. However, if there is no onset consonant, the speaker has not initiated to produce a fast constriction at the beginning. They have instead initiated to move the main mass of the tongue into the right position to produce the vowel. In EMA, the fast consonant movement registers early, but the slow vowel movement registers only after a delay when the muscles over come the inertia of the mass.

The rest position was not specified in the instructions of the experiments of this thesis. This may have had an effect on the outcomes of the experiments. There is a relevant cross-over effect in the articulatory onset data of Kawamoto et al. (2008) (see Section 2.2.2 for discussion of their experiments). They checked the starting positions of their participants from the video recordings and found that some consistently used an open-mouthed starting position, while others were consistent in starting a trial with their mouth closed, and a third group alternated between the two. Analysis of the consistent participants in the longer 600 ms delay condition, found that the open-mouthed participants initiated lip movement sooner for utterances with a bilabial onset, and later for an alveolar onset. An inverse of this pattern is evident in the data of the closed mouthed participants; they initiated lip movement later for a bilabial onset then for an alveolar onset. The acoustic onset times for both groups are very similar. This seems to provide evidence for a constant articulatory onset time with differing

articulatory initiation strategies depending on the starting position. However, this should be verified by studying this effect with data from other articulators. While such analysis is outwith the scope of this thesis, the data recorded here could potentially be used for the analysis by splining the UTI data and by identifying the starting positions in both UTI and EMA data recorded in this thesis.

7.2 Method development

The usefulness of the visualisation and automated analysis tools developed in this thesis is proven by their use in analysing data from the experiments. Basic PD provides an efficient way of evaluating the change present in a UTI video in a single glance. Combining PD visualisation with a basic annotation tool provides a fast way of extracting articulatory onset times from UTI data. Finally, the SBPD-based onset detection proved a valuable tool in analysing the extensive data sets of Experiments 2 and 3.

Following the example provided by the studies reviewed by Koppenhaver et al. (2009), PD results should be compared with tongue Electromyography (EMG). This would provide potentially explanations of the role of muscle activation in how articulation onsets behave in PD and more fine detail in how PD relates to articulatory gestures.

Further ways of analysing the PD contours should also be explored. Possible directions include at least applying Functional data analysis to collections of contour such as those in Figures 4.6 – 4.10, combining PD analysis with other movement analysis methods such as optic flow analysis and movement estimation based on extracted tongue surfaces, and further refinements of the PD metric itself by, for example, adding pre-processing stages such as edge preserving smoothing. Improving the data efficiency of automatic analysis of SBPD so that most of recorded tokens could be used in the analysis is another important future goal.

Since use of Matlab is not very wide spread among speech researchers, the

PD tools will be ported to a completely free open source platform in the near future. Computational performance will be the main deciding factor in whether this should be Python/SciPy/NumPy, Julia or some other platform.

7.3 General discussion

Sternberg et al. (1978; 1988) base their model of speech production on data sets that show a linearly increasing delay of the onset of acoustic speech as the number of words or stress units in the utterance increases. Their data also shows that articulation rate of the utterances decreases as the utterance length increases. Given the result on decreasing articulation rate increasing AAI, it seems likely that at least part of the linear increasing in acoustic latency in Sternberg et al.'s data is not due to increased latency, but in increased AAI. It is possible that even the whole latency effect in their data is only an articulation rate effect with no actual change in how long it takes to start the response regardless of the length of the list to be produced.

7.3.1 Speech ready position

Speech ready position has received attention in the literature over a long period of time. It and closely related concepts have been put forward as part of several theories and models. Especially in articulatory speech modelling, the speech ready position is a natural thing to postulate: a model has to start its movement somewhere – a model needs an origin state for the articulators.

Evidence of the existence of speech ready position as an independent target has been difficult to come by with some of the most convincing studies equating it with the inter-speech posture, but none that the author is aware of that would have shown how to reliably elicit it at the onset of an utterance.

At the beginning of this thesis project, studying the occurrence of the speech ready position as part of preparatory movements was one of the tentative goals. However, quite early on, informal analysis pointed strongly to speakers not moving from rest first to a speech ready position before initiating segmental

articulation, not even when given time to do so in a picture naming context.

Regardless of the above, absence of evidence is not evidence of absence. In theory, the speech ready position could be the initial target of articulation when initiating speech. If this was the case we might expect speakers to produce two gestures at the beginning of each utterance: First, a movement to a speech ready position, and second, the first gesture of the first target sound. There is no evidence in the data of this thesis, that this would be the case.

It should also be noted that Experiment 1 is a picture naming experiment, while Experiments 2 and 3 are delayed naming experiments. This means that in Experiment 1 the participant does not know beforehand what they are going to say, while in Experiments 2 and 3 they do. The major difference between the data from these two paradigms is, that there are hesitations and vocalisations not related to the target word present in the picture naming data, but not in the delayed naming data.

This leads to the conclusion that either speech ready position is not part of pre-speech articulation in the delayed naming experiments, or that the speaker is already at speech ready position before the recording begins. The latter sounds reasonable given that the speakers know that they are doing a speeded trial. However, without analysing the starting positions, this remains only a hypothesis.

7.3.2 Open questions and future work

In the list below summarises possible directions of future research in continuation of this thesis project.

- Was there change from one session to the next in the rest positions employed by the speakers?
- How does the timing of initiating longer/multi-syllable utterances pattern?
- How are different phonation initiation strategies conditioned by the onset sound quality and how do they reflect in, for example, /VC/ word timing

patterns?

- Is there so much variation and is that variation in some sense systematic in initial consonants (for example, in pre-aspiration of /w/ and full voicing of unaspirated plosives) that even results from matched onset studies become unreliable?
- The results on localised movement onsets illustrated in Figures 5.6 – 5.9, would benefit from more comprehensive sampling of the possible /CC/ and /CCC/ onset combinations. It would also be useful to have access to anatomical measurements of the participants in order to be able to check if individual anatomy affects the motor strategies or if they are more a result of individual habits.
- The long span of the recordings gives the opportunity to test whether each individual has their own habitual rest position and speech preparation pattern. However, since tongue surface contours have not yet been extracted from the data, analysing the rest positions remains future work.
- A related scientific application area of the tools developed here is analysis of discourse behaviours such as turn taking. From an articulatory point of view discourse behaviour involves alternating between speaking, listening, and negotiating turns. The latter involves some of the same processes as the ones that are present in pre-speech situations such as the picture naming task studied in this project. They also involve signalling the interlocutor that one would like to take a turn and/or that one is either ready to yield the turn or wants to keep speaking.

Bibliography and References

Abbs, J. H. and Gracco, V. L. (1983). Sensorimotor actions in the control of multi-movement speech gestures. *Trends in Neurosciences*, 6(9):391 – 395.

Articulate Instruments Ltd (2008). *Ultrasound Stabilisation Headset Users Manual: Revision 1.4*. Edinburgh, UK: Articulate Instruments Ltd.

Articulate Instruments Ltd (2012). *Articulate Assistant Advanced User Guide: Version 2.14*. Edinburgh, UK: Articulate Instruments Ltd.

Baayen, R. H. (2008). *Analyzing Linguistic Data, A Practical Introduction to Statistics using R*. Cambridge University Press, Cambridge.

Barbosa, A., Yehia, H., and Vatikiotis-Bateson, E. (2008). Linguistically valid movement behavior measured non-invasively. In Göcke, R., Lucey, P., and Lucey, S., editors, *Proceedings of the International Conference on Auditory-Visual Speech Processing*, pages 173 – 177.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1 – 48.

Bird, S., Moisik, S. R., Leonard, J., and Smith, S. (2010). An optic flow analysis of tongue movement in SENCOTEN /qV/ and /Vq/ sequences. In *Ultrafest V*, Ultrafest V, Haskins Lab, New Haven, Connecticut.

Birkholz, P. (2005). *3D-Artikulatorische Sprachsynthese*. PhD thesis, University of Rostock, Berlin. Logos Verlag.

Boersma, P. and Weenink, D. (2010). Praat: doing phonetics by computer [computer program]. Version 5.1.44, retrieved 4 October 2010 from `http://www.praat.org/`.

Bohland, J. W., Bullock, D., and Guenther, F. H. (2010). Neural representations and mechanisms for the performance of simple speech sequences. *Journal of Cognitive Neuroscience*, 22(7):1504 – 1529.

Branderud, P. (1985). Movetrack - a movement tracking system. In *Proceedings of the French-Swedish Symposium on Speech*, pages 113 – 122, Grenoble.

Browman, C. P. and Goldstein, L. (1988). Some notes on syllable structure in articulatory phonology. *Phonetica*, 45:140 – 155.

Browman, C. P. and Goldstein, L. (1990). Articulatory gestures as phonological units. *Phonology*, 6:201 – 251.

Browman, C. P. and Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, 49:155 – 180.

C-K., C. and Wang, W. S.-Y. (1978). Use of optical distance sensing to track tongue motion. *Journal of Speech and Hearing Research*, 21:482 – 496.

Carlsen, A. N., Maslovat, D., and Franks, I. M. (2012). Preparation for voluntary movement in healthy and clinical populations: evidence from startle. *Clinical Neurophysiology*, 123:21 – 33.

Carstens Medizinelektronik (2015). *Articulography - electromagnetic systems for visualization of speech movement inside the mouth*. Website `http://www.articulograph.de/` accessed 27 March 2015.

Cattell, J. M. (1886). The time taken up by cerebral operations. *Mind*, 11:220 – 242.

Chiba, T. and Kajiyama, M. (1941). *The Vowel, Its Nature and Structure*. Phonetic Society of Japan.

Chiu, C. and Gick, B. (2014). Startling speech: eliciting prepared speech using startling auditory stimulus. *Frontiers in Psychology*, 5(1082).

Cho, T. and Keating, P. (2009). Effects of initial position versus prominence in english. *Journal of Phonetics*, 37(4):466 – 485.

Cleland, J., Lloyd, S., Campbell, L., Crampin, L., Palo, P., Sugden, E., Wrench, A., and Zharkova, N. (2019). The impact of real-time articulatory information on phonetic transcription: Ultrasound-aided transcription in cleft lip and palate speech. *Folia Phoniatrica et Logopaedica*. accepted for publication.

Collins, B. and Mees, I. M. (1995). Approaches to articulatory setting in foreign-language teaching. In Lewis, J. W., editor, *Studies in General and English Phonetics: Essays in Honour of Professor J. D. O'Connor*, pages 415 – 424. Routledge, New York.

Csapó, T. G. and Lulich, S. M. (2015). Error analysis of extracted tongue contours from 2D ultrasound images. In *INTERSPEECH-2015*, pages 2157 – 2161.

Dalston, R. M. and Keefe, M. J. (1988). Digital, labial, and velopharyngeal reaction times in normal speakers. *Cleft Palate Journal*, 25(3):203 – 209.

Dart, S. (1987). A bibliography of x-ray studies of speech. *Working papers in Phonetics, UCLA Phonetics Laboratory Group*, 66:1 – 97.

De Boer, T. J. (2003). *History of Philosophy in Islam*. Kessinger Publishing.

Deary, I. J., Liewald, D., and Nissan, J. (2011). A free, easy-to-use, computer-based simple and four-choice reaction time programme: The deary-liewald reaction time task. *Behavior Research Methods*, 43(1):258 – 268.

Dedouch, K., Horáček, J., Vampola, T., and Černý, L. (2002). Finite element modelling of a male vocal tract with consideration of cleft palate. In *Forum Acusticum*, Sevilla, Spain.

Demolin, D., George, M., Lecuit, V., Metens, T., Soquet, A., and Raeymaekers, H. (1997). Coarticulation and articulatory compensations by dynamic mri. In *Proceedings of Eurospeech '97*, pages 43–46.

Demolin, D., Metens, T., and Soquet, A. (2000). Real time mri and articulatory coordinations in vowels. In *Proceedings of the 5th Speech Production Seminar: Models and Data*, pages 86 – 93, München, Germany.

Drake, E., Schaeffler, S., and Corley, M. (2013a). Articulatory evidence for the involvement of the speech production system in the generation of predictions during comprehension. In *Architectures and Mechanisms for Language Processing (AMLaP)*, Marseille.

Drake, E., Schaeffler, S., and Corley, M. (2013b). Does prediction in comprehension involve articulation? evidence from speech imaging. In *11th Symposium of Psycholinguistics (SCOPE)*, Tenerife.

Draper, M. H., Ladefoged, P., and Whitteridge, D. (1960). Expiratory pressures and air flow during speech. *British Medical Journal*, 1(5189):1837 – 1843.

Engwall, O. (2000a). A 3d tongue model based on mri data. In *In Proceedings of International Conference on Spoken Language Processing 2000 (ICSLP 2000)*, pages III: 901–904.

Engwall, O. (2000b). Dynamical aspects of coarticulation in swedish fricatives - a combined ema & epg study. *TMH-QPSR*, 4/2000:49–73.

Engwall, O. and Badin, P. (1999). Collecting and analysing two- and three-dimensional MRI data for swedish. *TMH-QPSR*, 3-4/1999:11–38.

Ericsdotter, C. (2005). *Articulatory-Acoustic Relationships in Swedish Vowel Sounds*. PhD thesis, Stockholm University, Stockholm, Sweden.

Euclid (2006). *The Elements: Books I-XIII - Complete and Unabridged*. Barnes & Noble. Translated by Sir Thomas Heath.

Fant, G. (1960). *Acoustic Theory of Speech Production*. Mouton, The Hague.

Farnetani, E. and Recasens, D. (1999). Coarticulation models in recent speech production theories. In Hardcastle, W. J. and Hewlett, N., editors, *Coarticulation: Theory, Data and Techniques*, pages 31 – 68. Cambridge University Press.

Fasel, I. and Berry, J. (2010). Deep belief networks for real-time extraction of tongue contours from ultrasound during speech. In *2010 20th International Conference on Pattern Recognition*, pages 1493 – 1496.

Feher, J. (2012). *Quantitative Human Physiology*. Academic Press.

Feldman, A. (1986). Once more on the equilibrium-point hypothesis (lambda model) for motor control. *Journal of Motor Behavior*, 18(1):17 – 56.

Fitt, S. (2014). Unisyn lexicon release. version 1.3, retrieved 26 November 2014 from `http://www.cstr.ed.ac.uk/projects/unisyn/`.

Fletcher, S. G., Dagenais, P. A., and Critz-Crosby, P. (1991). Teaching vowels to profoundly hearing-impaired speakers using glossometry. *Journal of Speech and Hearing Research*, 34:943 – 956.

Fougeron, C., D'Alessandro, D., and Lancia, L. (2018). Reduced coarticulation and aging. *The Journal of the Acoustical Society of America*, 144(3):1905–1905.

Fuchs, S. and Ünal-Logacev, O. (2017). Tongue palatal contacts during speech preparation in 7 languages. Poster presented at Speech Motor Control, Groningen, Netherlands.

Fujimura, O. (1991). Recording and interpreting articulatory data - microbeam and other methods. In *Proceedings of the XIIth ICPhS*, volume 3, pages 120–124.

Fuller, D. R., Pimentel, J. T., and Peregoy, B. M. (2012). *Applied anatomy & physiology for speech-language pathology & audiology*. Wolters Kluwer-Lippincott Williams & Wilkins, Baltimore, MD.

Gick, B., Wilson, I., and Derrick, D. (2013). *Articulatory Phonetics*. Wiley-Blackwell.

Gomi, H., Honda, M., Ito, T., and Murano, E. Z. (2002). Compensatory articulation during bilabial fricative production by regulating muscle stiffness. *Journal of Phonetics*, 30(3):261 – 279.

Granström, B. and House, D. (2007). Inside out - acoustic and visual aspects of verbal and non-verbal communication. In *Proceedings of ICPhS 2007*, pages 11 – 18.

Grimaldi, M., Fivela, B. G., Sigona, F., Tavella, M., Fitzpatrick, P., Craighero, L., Fadiga, L., Sandini, G., and Metta, G. (2008). New technologies for simultaneous acquisition of speech articulatory data: 3D articulograph, ultrasound and electroglottograph. In *LangTech*.

Guenther, F. H. (2016). *Neural Control of Speech*. The MIT Press, Cambridge, MA.

Guenther, F. H., Ghosh, S. S., and Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain & Language*, 96:280 – 301.

Hannukainen, A., Lukkari, T., Malinen, J., and Palo, P. (2007). Vowel formants from the wave equation. *Journal of the Acoustical Society of America Express Letters*, 122(1):EL1–EL7.

Harandi, N. M., Woo, J., Stone, M., Abugharbieh, R., and Fels, S. (2014). Subject-specific biomechanical modelling of the tongue: Analysis of muscle activations during speech. In *Proceedings of the 10th International Seminar on Speech Production*, Cologne.

Hardcastle, W., Gibbon, F., and Nicolaidis, K. (1991). Epg data reduction methods and their implications for studies of lingual coarticulation. *Journal of Phonetics*, 19:251–266.

Hardcastle, W. J. (1972). The use of electropalatography in phonetic research. *Phonetica*, 25:197 – 215.

Heffner, R.-M. S. (1950). *General phonetics*. The University of Wisconsin Press, Madison, WI.

Hewlett, N. and Beck, J. (2006). *An Introduction to the Science of Phonetics*. Routledge.

Hixon, T. J. (1971). An electromagnetic method for transducing jaw movements during speech. *Journal of the Acoustical Society of America*, 49:603 – 606.

Hoole, P. and Nguyen, N. (1999). Electromagnetic articulography in coarticulation research. In Hardcastle, W. J. and Hewlett, N., editors, *Coarticulation: Theory, Data and Techniques*, pages 260 – 269. Cambridge University Press.

Hoole, P. and Zierdt, A. (2010). Five-dimensional articulography. *Speech motor control*, pages 331–349.

Hoole, P., Zierdt, A., and Geng, C. (2003). Beyond 2d in articulatory data acquisition and analysis. In *The 15th International Congress of Phonetic Sciences*, pages 265 – 268.

Horbatiuk, I. (2011). Fourier & wavelet methods for finding speech onset latencies. Master's thesis, McMaster University.

Horn, B. K. P. and Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence*, 17:185 – 203.

Hueber, T., Aversano, G., Chollet, G., Denby, B., Dreyfus, G., Oussar, Y., Roussel, P., and Stone, M. (2007). Eigentongue feature extraction for an ultrasound-based silent speech interface. In *Proceedings of ICASSP 2007*, pages I: 1245 – 1248.

Indefrey, P. and Levelt, W. (2004). The spatial and temporal signatures of word production components. *Cognition*, 92(1):101 – 144.

Izdebski, K. and Shipp, T. (1978). Minimal reaction times for phonatory initiation. *Journal of Speech and Hearing Research*, 21(4):638 – 651.

Jannedy, S., Fuchs, S., and Weirich, M. (2010). Articulation beyond the usual: Evaluating the fastest german speaker under laboratory conditions. In Fuchs, S., Hoole, P., Mooshammer, C., and Zygis, M., editors, *Between the regular and the particular in speech and language*, pages 205 – 234. Peter Lang Verlag.

Jenner, B. (2001). 'Articulatory setting': genealogies of an idea. *Historiographica Linguistica*, pages 121 – 141.

Jensen, A. R. (2002). Galton's legacy to research on intelligence. *Journal of Biosocial Science*, 34(2):145 – 172.

Jescheniak, J. D. and Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20:824 – 843.

Kapusinski, D. A. and Rosenquist, H. S. (1973). A brief history of the voice key. In *Proceedings of the Annual Convention of the American Psychological Association*, volume 8, pages 945 – 946.

Kawamoto, A. H., Liu, Q., Mura, K., and Sanchez, A. (2008). Articulatory preparation in the delayed naming task. *Journal of Memory and Language*, 58(2):347 – 365.

Kelly, J. and Lochbaum, C. (1962). Speech synthesis. In *Proceedings of the 4th International Congress on Acoustics*, pages Paper G42: 1–4.

Kelz, H. P. (1971). Articulatory basis and second language learning. *Phonetica*, 24:193 – 211.

Kessler, B., Treiman, R., and Mullenix, J. (2002). Phonetic biases in voice key response time measurements. *Journal of Memory and Language*, 47:145 – 171.

Kiritani, S., Itoh, K., and Fujimura, O. (1975). Tongue-pellet tracking by a computer controlled x-ray microbeam system. *Journal of the Acoustical Society of America*, 48(6):1516–1520.

Kittredge, A. K., Dell, G. S., Verkuilen, J., and Schwartz, M. F. (1994). Where is the effect of frequency in word production? insights from aphasic picture-naming errors. *Cognitive neuropsychology*, 25(4):463 – 492.

Klapp, S. and Erwin, C. (1976). Relation between programming time and duration of the response being programmed. *Journal of experimental psychology. Human perception and performance*, 2(4):591 – 598.

Klapp, S. T., Anderson, W. G., and Berrian, R. W. (1973). Implicit speech in reading: Reconsidered. *Journal of Experimental Psychology*, 100(2):368 – 374.

Koppenhaver, S. L., Hebert, J. J., Parent, E. C., and Fritz, J. M. (2009). Rehabilitative ultrasound imaging is a valid measure of trunk muscle size and activation during most isometric sub-maximal contractions: a systematic review. *Australian Journal of Physiotherapy*, 55(3):153 – 169.

Krmpotić, J. (1958). Anatomisch-histologische und funktionelle verhältnisse des rechten und des linken nervus recurrens mit rücksicht auf die geschwindigkeit der impulsleitung bei einer ursprungsanomalie der rechten schlüsselbeinarterie. *Arch. Ohr. - Nas. - Kehlk. Heilk.*, 173:490 – 496.

Krmpotić, J. (1959). Donneés anatomique et histologiques relatives aux effecteurs laryngo-pharyngo-buccaux. *Rev. Laryngol.*, 11:829 – 848.

Kroos, C. (2012). Evaluation of the measurement precision in three-dimensional electromagnetic articulography (carstens AG500). *Journal of Phonetics*, 40:453 – 465.

Kühnert, B. and Nolan, F. (1999). The origin of coarticulation. In Hardcastle, W. J. and Hewlett, N., editors, *Coarticulation: Theory, Data and Techniques*, pages 7 – 30. Cambridge University Press.

Ladefoged, P. (1995). *A Course in Phonetics*. Harcourt Brace Jovanovich, Inc., Orlando, Florida.

Ladefoged, P., Anthony, J., and Riley, C. (1971). Direct measurement of the vocal tract. *Working papers in Phonetics, UCLA Phonetics Laboratory Group*, 19.

Lammert, A., Ramanarayanan, V., Proctor, M., and Narayanan, S. (2013). Vocal tract cross-distance estimation from real-time MRI using region-of-interest analysis. In *Interspeech 2013*, pages 959 – 962.

Lance, D. and van der Giet, G. (1974). A computer on-line method for measuring articulatory movements. In Fant, G., editor, *Speech Communication Seminar*, pages 73 – 77, Stockholm.

Laprie, Y., Vaxelaire, B., and Cadot, M. (2014). Geometric articulatory model adapted to the production of consonants. In *Proceedings of 10th ISSP*, pages 253 – 256.

Laver, J. (1978). The concept of articulatory settings: an historical survey. *Historiographica Linguistica*, V:1 – 14.

Lee, A. and Doherty, R. (2017). Speaking rate and articulation rateof native speakers of irish english. *Speech, Language and Hearing*, 20(4):206 – 211.

Lehiste, I. (1970). *Suprasegmentals*. M.I.T. Press, Cambridge, Massachusetts.

Lemmetty, S. (1999). Review of speech synthesis technology. Master's thesis, Helsinki University of Technology.

Lenneberg, E. H. (1967). *Biological Foundations of Language*. John Wiley & Sons, New York.

Levelt, W. J. M. (1989). *Speaking, From Intention to Articulation*. The MIT Press.

Levelt, W. J. M., Roelofs, A., and Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22:1 – 75.

Lim, Y., Lingala, S. G., Toutios, A., Narayanan, S., and Nayak, K. S. (2016). Improved depiction of tissue boundaries in vocal tract real-time mri using automatic off-resonance correction. In *Interspeech 2016*, pages 1765 – 1769, San Francisco, USA.

Lingala, S. G., Zhu, Y., Kim, Y.-C., Toutios, A., Narayanan, S., and Nayak, K. S. (2017). A fast and flexible MRI system for the study of dynamic vocal tract shaping. *Magnetic resonance in medicine*, 77 1:112–125.

Lloyd, J., Stavness, I., and Fels, S. (2011). The ArtiSynth Toolkit For Rigid-Deformable Biomechanics. In *ISB Technical Group on Computer Simulation Symposium, Poster*.

Lukkari, T., Malinen, J., and Palo, P. (2007). Recording speech during magnetic resonance imaging. In *MAVEBA 2007*, pages 163 – 166, Florence, Italy.

Maeda, S. (1982). A digital simulation method of the vocal-tract system. *Speech Communication*, 1:199 – 229.

Maeda, S. (1990). Compensatory articulation during speech: Evidencefrom the analysis and synthesis of vocal-tract shapes using an articulatory model. In Hardcastle, W. J. and Marchal, A., editors, *Speech Production and Speech Modelling*, pages 131 – 149. Boston: Kluwer Academic Publishers.

Malinen, J. and Palo, P. (2009). Recording speech during MRI: Part II. In *MAVEBA 2009*, pages 211–214, Florence, Italy.

Manuel, S. (1999). Cross-language studies: relating language-particular coarticulation patterns to other language-particular facts. In Hardcastle, W. J. and Hewlett, N., editors, *Coarticulation: Theory, Data and Techniques*, pages 179 – 198. Cambridge University Press.

McMillan, C. T. and Corley, M. (2010). Cascading influences on the production of speech: Evidence from articulation. *Cognition*, 117(3):243 – 260.

Mermelstein, P. (1973). Articulatory model for the study of speech production. *Journal of the Acoustical Society of America*, 53(4):1070 – 1082.

Meyer, B. U., Werhahn, K., Rothwell, J. C., Roericht, S., and Fauth, C. (1994). Functional organisation of corticonuclear pathways to motoneurones of lower facial muscles in man. *Experimental Brain Research*, 101:465 – 472.

Mitsuya, T., MacDonald, E. N., Purcell, D. W., and Munhall, K. G. (2011). A cross-language study of compensation in response to real-time formant perturbation. *The Journal of the Acoustical Society of America*, 130(5):2978 – 2986.

Moisik, S. R. (2010). Laryngeal ultrasound assessment of retracted and constricted articulations by phoneticians. In *Ultrafest V*, Ultrafest V, Haskins Lab, New Haven, Connecticut.

Mooshammer, C., Goldstein, L., Nam, H., McClure, S., Saltzman, E., and Tiede, M. (2012). Bridging planning and execution: Temporal planning of syllables. *Journal of Phonetics*, 40:374 – 389.

Müller, M. (2007). *Information Retrieval for Music and Motion*. Springer, Berlin Heidelberg New York.

Munhall, K., Vatikiotis-Bateson, E., and Tokhura, Y. (1995). X-ray film database for speech research. *Journal of the Acoustical Society of America*, 98(2):1222–1224.

Neurobehavioral systems (2015). *Presentation®*. Neurobehavioral systems. Version 14.2.

Nieto-Castanon, A., Guenther, F. H., Perkell, J. S., and Curtin, H. D. (2005). A modeling investigation of articulatory variability and acoustic stability during American English /r/ production. *The Journal of the Acoustical Society of America*, 117(5):3196 – 3212.

Niziolek, C. A. and Guenther, F. H. (2013). Vowel category boundaries enhance cortical and behavioral responses to speech feedback alterations. *Journal of Neuroscience*, 33(29):12090 – 12098.

O'Connor, J. D. (1973). *Phonetics*. Penguin, Harmondsworth.

Oller, D. K. (1973). The effect of position in utterance on speech segment duration in english. *The Journal of the Acoustical Society of America*, 54(5):1235 – 1247.

Palo, P. (2011). *A wave equation model for vowels: Measurements for validation*. Licentiate thesis, Institute of Mathematics, Aalto University, Helsinki.

Perkell, J. and Oka, D. (1980). Use of an alternating magnetic field device to track midsagittal plane movements of multiple points inside the vocal tract. *Journal of the Acoustical Society of America*, 67:92.

Possamaï, C., Burle, B., Osma, A., and Hasbroucq, T. (2002). Partial advance information, number of alternatives, and motor processes: an electromyographic study. *Acta Psychologica*, 111(1):125 – 139.

Preston, J. L., McAllister Byun, T., Boyce, S. E., Hamilton, S., Tiede, M., Phillips, E., Rivera-Campos, A., and Whalen, D. H. (2017). Ultrasound images of the tongue: A tutorial for assessment and remediation of speech sound errors. *Journal of visualized experiments : JoVE*, 119:55123.

Purcell, D. W. and Munhall, K. G. (2006). Compensation following real-time manipulation of formants in isolated vowels. *The Journal of the Acoustical Society of America*, 119(4):2288 – 2297.

Python Software Foundation (2017). *Python Language Reference, version 2.7*. Python Software Foundation.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raeesy, Z., Baghai-Ravary, L., and Coleman, J. (2011). Parametrising degree of articulator movement from dynamic MRI data. In *12th Interspeech*, pages 2853 – 2856.

Rastle, K. and Davis, M. H. (2002). On the complexities of measuring naming. *Journal of Experimental Psychology: Human Perception and Performance*, 28(2):307 – 314.

Rastle, K., Harrington, J. M., Croot, K. P., and Coltheart, M. (2005). Characterizing the motor execution stage of speech production: Consonantal effects on delayed naming latency and onset duration. *Journal of Experimental Psychology: Human Perception and Performance*, 31(5):1083 – 1095.

Raudies, F. (2013). Optic flow. *Scholarpedia*, 8(7):30724. `http://www.scholarpedia.org/article/Optic_flow`.

Riès, S., Legou, T., Burle, B., Alario, F. X., and Malfait, N. (2012). Why does picture naming take longer than word reading? the contribution of articulatory processes. *Psychonomic Bulletin & Review*, 19(5):955 – 961.

Riès, S., Legou, T., Burle, B., Alario, F. X., and Malfait, N. (2014). Corrigendum to "why does picture naming take longer than word reading? the contribution of articulatory processes". *Psychonomic Bulletin & Review*, pages 1 – 3.

Roon, K. D. (2013). *The dynamics of phonological planning*. PhD thesis, New York University.

Rosenfelder, I., Fruehwald, J., Evanini, K., and Yuan, J. (2011). Fave (forced alignment and vowel extraction) program suite. http://fave.ling.upenn.edu.

Rossing, T. D., Moore, F. R., and Wheeler, P. A. (2002). *The Science of Sound*. Addison Wesley, San Francisco, California, 3rd edition.

Rossion, B. and Pourtois, G. (2004). Revisiting snodgrass and vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception*, 33:217 – 236.

Saltzman, E. L. and Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1(4):333 – 382.

Savariaux, C., Badin, P. andSamson, A., and Gerbera, S. (2017). A comparative study of the precision of carstens and northern digital instruments electromagnetic articulographs. *The Journal of the Acoustical Society of America*, 60(2):322 – 340.

Schaeffler, S., Scobbie, J., and Schaeffler, F. (2014). Measuring reaction times: Vocalisation vs. articulation. In *Proceedings of 10th ISSP*, pages 383 – 386.

Schaeffler, S., Scobbie, J., and Schaeffler, F. (2015). Complex patterns in silent speech preparation: Preparing for fast response might be different to preparing for fast speech in a reaction time experiment. In *Proceedings of ICPhS 2015*, Glasgow, UK.

Schaeffler, S., Scobbie, J. M., and Mennen, I. (2008). An evaluation of inter-speech postures for the study of language-specific articulatory settings. In *8th International Seminar on Speech Production (ISSP 2008)*, pages 121 – 124.

Schönle, P., Gräbe, K., Wenig, P., Höhne, J., Schrader, J., and Conrad, B. (1987). Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. *Brain and Language*, 31:26 – 35.

Schroeder, C. E. and Foxe, J. J. (2002). The timing and laminar profile of converging inputs to multisensory areas of the macaque neocortex. *Cognitive Brain Research*, 14:187 – 198.

Scobbie, J. M., Gordeeva, O. B., and Matthews, B. (2007). Scottish english speech acquisition. In McLeod, S., editor, *The International Guide to Speech Acquisition*, pages 221 – 240. Thomson Delmar Learning, Clifton Park, NY.

Scobbie, J. M., Lawson, E., Cowen, S., Cleland, J., and Wrench, A. A. (2011). A common co-ordinate system for mid-sagittal articulatory measurement. Technical report, Queen Margaret University.

Siegenthaler, B. M. and Hochberg, I. (1965). Reaction time of the tongue to auditory and tactile stimulation. *Perceptual and Motor Skills*, 21:387 – 393.

Simko, J. (2009). *The Embodied Modelling of Gestural Sequencing in Speech*. PhD thesis, UCD School of Computer Science and Informatics.

Snodgrass, J. G. and Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, 6(2):174 – 215.

Sonoda, Y. (1974). Observation of tongue movements employing magnetometer sensor. *IEEE Trans. Magn.*, 10:954–957.

Stavness, I., Lloyd, J. E., Payan, Y., and Fels, S. (2011). Coupled hard-soft tissue simulation with contact and constraints applied to jaw-tongue-hyoid dynamics. *INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN BIOMEDICAL ENGINEERING*, 27(3):367 – 390.

Sternberg, S., Knoll, R. L., Monsell, S., and Wright, C. E. (1988). Motor programs and hierarchical organization in the control of rapid speech. *Phonetica*, 45:175 – 197.

Sternberg, S., Monsell, S., Knoll, R. L., and Wright, C. E. (1978). The latency and duration of rapid movement sequences: Comparisons of speech and typewriting. In Stelmach, G. E., editor, *Information Processing in Motor Control and Learning*. Academic Press.

Stevens, K. N. (1998). *Acoustic Phonetics*. The MIT Press, Cambridge, Massachusetts.

Stockard, J. J., Stockard, J. E., and Sharbrough, F. W. (1977). Detection and localization of occult lesions with brain-stem auditory responses. *Mayo Clinic Proceedings*, 52:761 – 769.

Stone, M. (2005). A guide to analyzing tongue motion from ultrasound images. *Clinical Linguistics and Phonetics*, 19(6–7):455 – 502.

Stone, S. and Birkholz, P. (2017). Angle correction in optopalatographic tongue distance measurements. *IEEE Sensors Journal*, 17(2):459 – 468.

Sumby, W. H. and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, 26(2):212 – 215.

Suomi, K., Toivanen, J., and Ylitalo, R. (2008). *Finnish Sound Structure – Phonetics, phonology, phonotactics and prosody*. STUDIA HUMANIORA OULUENSIA. University of Oulu.

Titze, I. (1980). Comments on the myoelastic - aerodynamic theory of phonation. *Journal of Speech and Hearing Research*, 23(3):495 – 510.

Titze, I. R. (2008). Nonlinear source-filter coupling in phonation: Theory. *Journal of the Acoustical Society of America*, 123(5):2733 – 2749.

Turk, A., Nakai, S., and Sugahara, M. (2006). Acoustic segment durations in prosodic research: a practical guide. In Sudhoff, S., Lenertová, D., Meyer, R., Pappert, S., Augurzky, P., Mleinek, I., Richter, N., and Schliesser, J., editors, *Methods in Empirical Prosody Research*, pages 1 – 28. De Gruyter, Berlin, New York.

Turk, A., Scobbie, J., Geng, C., Macmartin, C., Bard, E., Campbell, B., Dickie, C., Dubourg, E., Hardcastle, B., Hoole, P., Kanaida, E., Lickley, R., Nakai, S., Pouplier, M., King, S., Renals, S., Richmond, K., Schaeffler, S., Wiegand, R., White, K., and Wrench, A. (2010). The Edinburgh Speech Production Facility's articulatory corpus of spontaneous dialogue. *The Journal of the Acoustical Society of America*, 128(4):2429 – 2429.

Tyler, M. D., Tyler, L., and Burnham, D. K. (2005). The delayed trigger voice key: An improved analogue voice key for psycholinguistic research. *Behavior Research Methods*, 37(1):139 – 147.

Vampola, T., Laukkanen, A.-M., Horácek, J., and Švec, J. G. (2011). Vocal tract changes caused by phonation into a tube: A case study using computer tomography and finite-element modeling. *The Journal of the Acoustical Society of America*, 129(1):310 – 315.

Van Buuren, L. (1995). Postura: Clear and dark consonants, etcetera. In Lewis, J. W., editor, *Studies in General and English Phonetics: Essays in Honour of Professor J. D. O'Connor*, pages 130 – 142. Routledge, New York.

van der Linden, L., Riés, S., Legou, T., Burle, B., Malfait, N., and Alario, F.-X. (2014). On the fractionation of verbal response times. Poster presented at the International Workshop of Language Production. Geneva, Switzerland.

Vasseljen, O., Fladmark, A. M., Westad, C., and Torp, H. G. (2009). Onset in abdominal muscles recorded simultaneously by ultrasound imaging and intramuscular electromyography. *Journal of Electromyography and Kinesiology*, 19(2):e23 – e31.

Vogt, F., Lloyd, J. E., Perrier, P., Chabanas, M. Payan, Y., and Fels, S. (2006). An efficient biomechanical tongue model for speech research. In *In Proc. 7th International Seminar on Speech Production (ISSP)*, pages 51 – 58.

Wallis, J. (1653). *Grammatica linguae agnlicanae*. London: Longman. Edited and translated by J.A. Kemp 1972, originally published 1653.

Whalen, D., Iskarous, K., Tiede, M. K., Ostry, D. J., Lehnert-Lehouillier, H., Vatikiotis-Bateson, E., and Hailey, D. S. (2005). The haskins optically corrected ultrasound system (HOCUS). *Journal of Speech, Language, and Hearing Research*, 48:543 – 553.

Wilson, I. (2006). *Articulatory Settings of French and English monolingual and bilingual speakers*. PhD thesis, University of British Colombia.

Wrench, A., McIntosh, A., and Hardcastle, W. (1996). Optopalatograph (opg): A new apparatus for speech production analysis. In *Proceedings of the 4th ICSLP*, pages 1589–1592.

Wrench, A., McIntosh, A., and Hardcastle, W. (1997). Optopalatograph: development of a device for measuring tongue movement in 3d. In *Proceedings of Eurospeech '97*, pages 1055–1058.

Wrench, A., McIntosh, A., Watson, C., and Hardcastle, W. (1998). Optopalatograph: real-time feedback of tongue movement in 3d. In *Proceedings of the 5th ICSLP*, pages 305 – 308.

Wrench, A. A. and Scobbie, J. M. (2006). Spatio-temporal inaccuracies of video-based ultrasound images of the tongue. In *Proceedings of the 7th International Seminar on Speech Production*, pages 451 – 458, Ubatuba, Brazil.

Wrench, A. A. and Scobbie, J. M. (2011). Very high frame rate ultrasound tongue imaging. In *Proceedings of ISSP 9*, pages 155 – 162, Montreal.

Wrench, A. A. and Scobbie, J. M. (2016). Queen Margaret University ultrasound, audio and video multichannel recording facility (2008-2016). Technical report, Queen Margaret University.

Xu, K., Csapo, T. G., Roussel, P., and Denby, B. (2016). A comparative study on the contour tracking algorithms in ultrasound tongue images with automatic re-initialization. *Journal of the Acoustical Society of America Express Letters*, 139:EL1154.

Yamada, J. and Tamaoka, K. (2003). Measurement errors in voice-key naming latency for hiragana. *Perceptual and motor skills*, 97(3f):1100 – 1106.

Yunusova, Y., Green, J., and Mefferd, A. (2009). Accuracy assessment for AG500, electromagnetic articulograph. *Journal of Speech, Language and Hearing Research*, 52:547–555.

Zierdt, A., Hoole, P., Honda, M., Kaburagi, T., and Tillman, H. (2000). Extracting tongues from moving heads. In *Proceedings of the 5th Speech Production Seminar: Models and Data*, pages 313–316, München, Germany.

Appendix A

Forms and instructions used in Experiment 2

The forms and instruction sheets used in Experiment 2 can be downloaded from `http://taurlin.org/?page_id=126`.

Appendix B

Data processing tools

Matlab implementations of the PD and SBPD algorithms, the PD annotation tool, and Python implementation of the go-signal/beep detection algorithm can be downloaded from `http://taurlin.org/?page_id=126`.